**Name: SRI HARSHA**

**Senior GCP Data Engineer**

**Email Id:** [**harshab5096@gmail.com**](mailto:harshab5096@gmail.com)

**Contact Details: 980-946-5452**

**Professional Summary:**

* Over 9+ of IT experience, I have specialized as a Senior GCP Data Engineer, excelling in the design, development, maintenance, and support of Big Data applications. My expertise lies in Data Engineering, Data Pipeline Design, Development, and Implementation, serving in roles as a Data Engineer, Data Developer, and Data Modeler.
* Hands-on experience encompasses the migration of on-premise ETL processes to the Google Cloud Platform (GCP) using cloud-native tools such as BigQuery, Cloud Dataproc, Google Cloud Storage, and Composer. This experience is predominantly focused on IT data analytics projects.
* Have successfully implemented a wide array of solutions, including Big Data Analytics, Cloud Data Engineering, Data Warehousing/Data Mart, Data Visualization, Reporting, Data Quality, and Data Virtualization. Additionally, I am well-versed in AWS cloud services like EC2, S3, Glue, Athena, DynamoDB, and RedShift.
* Skill set extends to designing and deploying data engineering pipelines and performing data analysis using AWS services such as EMR, Glue, EC2, Lambda, Beanstalk, Athena, Redshift, Scoop, and Hive.
* Proficient in programming using Python, Scala, Java, and SQL, I have implemented Spark in EMR to process Enterprise Data across Data Lakes within AWS.
* Have designed and executed end-to-end data pipelines for the extraction, cleansing, processing, and analysis of substantial volumes of behavioral and log data, with a primary focus on data analytics in the AWS Cloud.
* Have contributed to infrastructure automation by writing Terraform scripts for AWS services, including ELB, CloudFront distribution, RDS, EC2, database security groups, Route 53, VPC, Subnets, Security Groups, and S3 Bucket. I have also converted existing AWS infrastructure to AWS Lambda deployed via Terraform and AWS CloudFormation.
* Hands-on experience extends to various GCP services, including BigQuery, Cloud Storage (GCS), Cloud Functions, Cloud Dataflow, Pub/Sub, Cloud Shell, GSUTIL, Data Proc, and Operations Suite (Stackdriver).
* Possess an in-depth understanding of SparkStreaming, SparkSQL, and other components of Spark, along with extensive experience in Scala, Java, Python, SQL, T-SQL, and R programming.
* Have developed and deployed enterprise-level applications using Hadoop ecosystem components like MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, and Kafka.
* Proficiency includes Spark Core, Spark SQL, Spark MLlib, Spark GraphX, and Spark Streaming for processing and transforming complex data using in-memory computing capabilities in Scala.
* Have integrated diverse data sources, including Oracle SE2, SQL Server, Flat Files, and Unstructured files, into data warehousing systems.
* Have extensive experience in developing MapReduce and Streaming jobs using Scala and Java for data cleansing, filtering, and aggregation, along with a comprehensive knowledge of the MapReduce framework.
* Well-versed in using IDEs such as Eclipse, IntelliJ IDE, PyCharm IDE, Notepad++, and Visual Studio for development.
* Background extends to Machine Learning algorithms and Predictive Modeling, covering Linear Regression, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, KNN, Neural Networks, and K-means Clustering, with a deep understanding of data architecture.
* Have experience working with NoSQL databases like Cassandra and HBase, enabling real-time read/write access to large datasets via HBase.
* Have developed Spark Applications capable of handling data from various RDBMS sources such as MySQL and Oracle Database, as well as streaming sources.
* Proficient in using GitHub/Git 2.12 for source and version control, I have a strong foundation in core Java concepts, including Object-Oriented Design (OOD) and Java components such as the Collections Framework, Exception handling, and I/O systems.

**Technical Skills:**

|  |  |
| --- | --- |
| **Languages** | Java, Scala, Python, SQL, and C/C++ |
| **Big Data Ecosystem** | Hadoop, MapReduce, Kafka, Spark, Pig, Hive, YARN, Flume, Sqoop,  Oozie, Zookeeper, Talend |
| **Hadoop Distribution** | Cloudera Enterprise, Data Bricks, Horton Works, EMC Pivotal. |
| **Databases** | Oracle, SQL Server, PostgreSQL |
| **Web Technologies** | HTML, CSS, JSON, JavaScript, Ajax |
| **Streaming Tools** | Kafka, RabbitMQ |
| **Cloud** | GCP, AWS, Azure, AWS EMR, Glue, RDS, Kinesis, DynamoDB, Redshift Cluster |
| **GCP Cloud** | Big Query, Cloud Data Proc, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Shell, GSUTIL, BQ Command Line, Cloud Data Flow |
| **AWS Cloud** | Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon Sage maker, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon Cloud Formation |
| **Azure Cloud** | Azure Data Lake, Data Factory, Azure SQL Database, Azure data bricks. |
| **Operating Systems** | Linux Red Hat/Ubuntu/CentOS, Windows 10/8.1/7/XP |
| **Testing** | Hadoop Testing, Hive Testing, MRUnit |
| **Application Servers** | Apache Tomcat, JBOSS, WebSphere |
| **Tools and Technologies** | Servlets, JSP, Spring (Boot, MVC, Batch, Security), Web Services,  Hibernate, Maven, GitHub, Bamboo. |
| **IDE’s** | IntelliJ, Eclipse, Net Beans. |

**Professional Experience:**

**Client: Chewy Dania Beach, FL April 2022 to Present**

**Role: Senior GCP Data Engineer**

**Responsibilities:**  
Collaborated with product teams to develop a wide range of store-level metrics and established data pipelines using Google Cloud Platform's (GCP) big data stack.

* Partnered with App teams to gather insights from Google Analytics 360 and constructed data marts in BigQuery to support analytical reporting for the sales and product teams.
* Proficient in GCP services such as Dataproc, Dataflow, PubSub, Google Cloud Storage (GCS), Cloud Functions, BigQuery, Stackdriver, Cloud Logging, IAM, and Data Studio for reporting.
* Developed a Python program with the Apache Beam framework and executed it in Cloud Dataflow to stream Pub/Sub messages into BigQuery tables.
* Successfully deployed streaming Maven build Cloud Dataflow jobs.
* Extensive experience in identifying and resolving production bugs in data using Stackdriver logs in GCP.
* Proficient in GCP services including Dataproc, GCS, Cloud Functions, Cloud SQL, and BigQuery.
* Utilized Cloud Shell SDK in GCP to configure services such as Dataproc, Storage, and BigQuery.
* Implemented partitioning and clustering for high-volume tables in BigQuery, optimizing query performance by leveraging high cardinality fields.
* Utilized Cloud Functions to facilitate data migration from BigQuery to downstream applications.
* Developed scripts using PySpark to push data from GCP to third-party vendors through their API framework.
* Proficient in building data pipelines within Airflow as a service (Composer) using various operators.
* Developed a Python program with Apache Beam to run data validation jobs between raw source files and BigQuery tables, executing it in Cloud Dataflow.
* Extensively employed Cloud Shell SDK in GCP for configuring and deploying services such as Cloud Dataproc (Managed Hadoop) and Google Storage on a daily basis.
* Created BigQuery jobs to load data into BigQuery tables from data files stored in Google Cloud Storage daily.
* Designed Tableau reports to track the dashboards published to Tableau Server, facilitating the identification of potential future clients in the organization.
* Assisted teams in identifying usage patterns of BigQuery and optimized BigQuery queries executed from Dataflow jobs, particularly concerning the utilization of BigQuery tables for store-level attributes.
* Loaded data every 15 minutes on an incremental basis into BigQuery, Google Dataproc, GCS buckets, Hive, Spark, Scala, Python, gsutil, and Shell Script.

**Environment:** GCP, Big Query, GCS, Python, Airflow, Data Procs, Shell Scripts, Scala, Spark, GCS, Pub/Sub, PySpark, Hadoop.

**Client: Merck Pharma, Branchburg, NJ September 2019 to March 2022**

**Role: GCP Data Engineer**

**Responsibilities:**

* Took part in the migration process of transitioning from an on-premises Hadoop system to utilizing Google Cloud Platform (GCP).
* Crafted scripts in Hive SQL and Presto SQL, leveraging Python plugins for both Spark and Presto to design intricate tables with optimal performance metrics such as partitioning, clustering, and skewing.
* Successfully migrated previously established cron jobs to Airflow/Composer within GCP.
* Utilized Confluence and Jira for project collaboration and management.
* Devised and implemented a configurable data delivery pipeline for scheduled updates to customer-facing data stores, constructed with Python.
* Possess proficiency in Machine Learning techniques, including Decision Trees, and Linear/Logistic Regressors, along with Statistical Modeling.
* Developed and deployed outcomes using Spark and Scala code within a Hadoop cluster running on GCP.
* Aggregated data from various sources to conduct intricate analyses, producing actionable results.
* Ensured the efficiency of the Hadoop/Hive environment while meeting Service Level Agreements (SLAs).
* Optimized the Tensorflow Model for enhanced efficiency.
* Designed a near-real-time data pipeline utilizing Flume, Kafka, and Spark Streaming to ingest client data from their web log server and apply necessary transformations.
* Conducted system analysis for identifying new enhancements/functionalities and performed impact analysis of the application for implementing ETL changes.
* Constructed high-performance, scalable ETL processes for loading, cleansing, and validating data.
* Developed Python Directed Acyclic Graphs (DAGs) in Airflow to orchestrate end-to-end data pipelines for multiple applications.
* Created BigQuery authorized views to enable row-level security and share data with other teams.
* Participated in the setup of the Apache Airflow service within GCP.
* Constructed data pipelines in Airflow within GCP for ETL-related tasks using various Airflow operators.

**Environment:** GCP, Biq Query,ETL,Airflow, Apache, Python, Hadoop, SQL, Spark, Data Procs, Cloud Composer, SQL Server.

**Client: Fifth Third Bank, Evansville, IN May 2017 to August 2019**

**Role: AWS Data Engineer**

**Responsibilities:**

* Designed and established an Enterprise Data Lake to support a variety of use cases, including analytics, data processing, storage, and reporting for vast and rapidly changing data.
* Took responsibility for maintaining high-quality reference data at its source by conducting operations such as data cleaning, transformation, and ensuring data integrity within a relational environment, working closely with stakeholders and solution architects.
* Architected and developed a Security Framework to provide fine-grained access to objects within AWS S3 using AWS Lambda and DynamoDB.
* Set up and managed Kerberos authentication principles to establish secure network communication within the cluster. Conducted testing for HDFS, Hive, Pig, and MapReduce to enable access for new users.
* Conducted end-to-end architecture and implementation assessments of various AWS services, including Amazon EMR, Redshift, and S3.
* Implemented machine learning algorithms using Python to predict the quantity a user might desire to order for a specific item, enabling automated suggestions via Kinesis Firehose and the S3 data lake.
* Utilized AWS EMR to transform and transfer large volumes of data to and from other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
* Employed Spark SQL with Scala and Python interfaces, automatically converting RDD case classes into schema RDDs.
* Imported data from various sources, such as HDFS and HBase, into Spark RDDs, and conducted computations using PySpark to generate the desired output responses.
* Developed Lambda functions with Boto3 to deregister unused Amazon Machine Images (AMIs) across all application regions, reducing costs associated with EC2 resources.
* Coded Teradata BTEQ scripts for data loading and transformation, addressing issues like SCD 2 date chaining and cleaning up duplicates.
* Created a reusable framework for future migrations, automating ETL processes from RDBMS systems to the Data Lake, utilizing Spark Data Sources and Hive data objects.
* Conducted data blending and preparation using Alteryx and SQL for Tableau consumption, publishing data sources to the Tableau server.
* Developed Kibana Dashboards based on Logstash data and integrated various source and target systems into Elasticsearch for near-real-time log analysis of end-to-end transactions.
* Implemented AWS Step Functions to automate and orchestrate tasks related to Amazon SageMaker, including publishing data to S3, training ML models, and deploying them for predictions.
* Integrated Apache Airflow with AWS to monitor multi-stage ML workflows, with tasks running on Amazon SageMaker.

**Environment:** AWS EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Amazon SageMaker, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Snowflake, Apache Pig, Python, SSRS, Tableau.

**Client: Grapesoft Solutions Hyderabad, India October 2015 to February 2017**

**Role: Azure Data Engineer**

**Responsibilities:**

* Utilized Azure Data Factory to seamlessly integrate data from both on-premises sources (including MySQL and Cassandra) and cloud-based platforms (such as Blob storage and Azure SQL Database), implementing necessary transformations before loading the data into Azure Synapse.
* Managed and configured resources efficiently across the cluster by leveraging Azure Kubernetes Service, while also scheduling their operations.
* Monitored the Spark cluster by employing Log Analytics and the Ambari Web UI. Transitioned log storage from Cassandra to Azure SQL Data Warehouse, resulting in enhanced query performance.
* Contributed to the development of data ingestion pipelines on the Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL. Additionally, worked with Cosmos DB, utilizing both the SQL API and Mongo API.
* Designed and constructed dashboards and visualizations to assist business users in data analysis, providing valuable insights to upper management, with a particular focus on Microsoft products such as SQL Server Reporting Services (SSRS) and Power BI.
* Led the migration of substantial datasets to Databricks (Spark), including the creation and administration of clusters, data loading, configuration of data pipelines, and loading data from Azure Data Lake Storage Gen2 to Databricks using ADF pipelines.
* Created various pipelines for data loading, beginning with Azure Data Lake and ending with Staging SQL Database, followed by loading the data into Azure SQL Database.
* Developed Databricks notebooks to streamline and curate data for various business use cases, also incorporating the mounting of blob storage within Databricks.
* Utilized Azure Logic Apps to construct workflows that scheduled and automated batch jobs, integrating various applications, ADF pipelines, and other services like HTTP requests and email triggers.
* Extensively worked with Azure Data Factory, focusing on data transformations, Integration Runtimes, Azure Key Vaults, triggers, and the migration of data factory pipelines to higher environments using ARM Templates.
* Ingested data in mini-batches and performed RDD transformations on these mini-batches using Spark Streaming for real-time streaming analytics within Databricks.

**Environment:** Azure SQL DW, Databrick, Azure Synapse, Cosmos DB, ADF, SSRS, Power BI, Azure Data lake, ARM, Azure HD Insight, Blob storage, Apache Spark.

**Client: Cybage Software Private Limited Hyd India May 2014 to September 2015**

**Role: Python Developer**

**Responsibilities:**

* Adopted a test-driven approach for application development and implemented unit tests using Python's Unit Test framework.
* Successfully migrated the Django database from SQLite to MySQL and PostgreSQL, ensuring complete data integrity throughout the process.
* Engaged in report creation using SQL Server Reporting Services (SSRS), developing various report types, including tables, matrices, and charts, and enabling web reporting by customizing URL Access.
* Crafted views and templates using Python and Django's view controller and templating language, resulting in a user-friendly website interface.
* Conducted API testing using the POSTMAN tool, employing various request methods such as GET, POST, PUT, and DELETE on each URL to verify responses and error handling.
* Designed Python and Bash tools to enhance the efficiency of the retail management application system and operations. These tools encompassed data conversion scripts, AMQP/Rabbit MQ, REST, JSON, and CRUD scripts for API integration.
* Carried out debugging and troubleshooting of web applications using Git as a version control tool, facilitating collaboration and coordination with team members.
* Developed and executed various MySQL database queries from Python, utilizing Python's MySQL connector and MySQL database package.
* Designed and maintained databases using Python and constructed Python-based APIs (RESTful Web Services) using SQL Alchemy and Postgre SQL.
* Created a web application using Python scripting for data processing, incorporating MySQL for database management, and using HTML, CSS, jQuery, and High Charts for data visualization on served pages.
* Dynamically generated property lists for each application, leveraging Python modules such as math, glob, random, itertools, functools, NumPy, matplotlib, seaborn, and pandas.
* Enhanced user experience by implementing navigation, pagination, column filtering, and the ability to add or remove desired columns for viewing, employing Python-based GUI components.

**Environment:** SQLite, MySQL, Postgre SQL, Python, Git, CRUD, POSTMAN, RESTful web service, SOAP, HTML, Git, CSS, jQuery, Django 1.4.

**Education:**

• Bachelor’s degree in computer science from Vellore Institute of Technology, Chennai- India from 2010 to 2014.