**Sai Kiran D**

**Sr. Data Engineer**

**Email:** **dsaikirandataengineer@gmail.com** **Phone no: +1 (312)-880-9385**

**Professional Summary:**

* Having 9+ years of professional IT experience which includes in Big Data ecosystem and Python technologies.
* Expertise in Hadoop architecture and various components such as HDFS, YARN, Hive, Pig, High Availability, Job Tracker, Task Tracker, Name Node, Data Node, Apache, Cassandra, and MapReduce programming paradigm.
* Experience in Hadoop cluster using Cloudera’s CDH, Hortonworks HDP.
* Worked on Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and am familiar with building custom Airflow operators and orchestrating workflows with dependencies involving multi-clouds.
* Experience in using various tools like Sqoop, Flume, Kafka, and Pig to ingest structured, semi-structured, and unstructured data into the cluster.
* Having hands-on experience in versioning using bit bucket.
* Good understanding of Cloud Based technologies such as AWS, and Azure.
* Proficient with Apache Spark ecosystems such as Spark and Spark Streaming using Scala and Python.
* Developed highly optimized Spark applications to perform various data cleansing, validation, transformation, and summarization activities according to the requirement.
* Hands-on Experience in Spark architecture and its integrations like Spark SQL, Data Frames, and Datasets APIs.
* Extensive experience in migrating on-premises Hadoop platforms to cloud solutions using AWS and GCP.
* Profound knowledge of Root Cause Analysis, Metadata Analysis, Tableau Dashboard, Tableau Calculated Fields, SQL, and Tableau Blend for code automation, customized reports, code re-usability, and Ad-hoc reports.
* Experience in RDD architecture and implementing spark operations on RDD also optimizing transformations and actions in Spark.
* Hands-on experience working on statistical procedures in R like ggplot, shiny apps, and dashboard building.
* Expertise in creating and customizing Splunk applications, searches, and dashboards as desired by IT teams and businesses.
* Experience collecting log data from various sources and integrating it into HDFS using Flume staging data in HDFS for further analysis.
* Designed and developed logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema and Slowly Changing Dimensions.
* Experience in Infrastructure Development and Operations involving AWS Cloud Services, EC2, EBS, VPC, RDS, SES, ELB, Auto scaling, CloudFront, Cloud Formation, Elastic Cache, API Gateway, Route 53, Cloud Watch, SNS.
* Experience in changing over existing AWS infrastructure to Serverless architecture (AWS Lambda, AWS Kinesis) through the creation of a Serverless Architecture using AWS Lambda, API gateway, Route 53, S3 buckets.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)
* Excellent experience in writing python as an ETL framework and Pyspark to process huge amounts of data daily.
* Strong experience in implementing data models and loading unstructured data using HBase, Dynamo DB, and Cassandra.
* Having hands-on experience in Application Deployment using the CICD pipeline.
* Can work parallel in both GCP and Aws Clouds coherently.
* Experience in implementing Spark using Scala and Spark SQL for faster processing of data.
* Strong experience in extracting and loading data using complex business logic using Hive from different data sources and building ETL pipelines to process terabytes of data daily.
* Experienced in transporting and processing real-time event streaming using Kafka and Spark Streaming.
* Hands-on experience with importing and exporting data from Relational databases to HDFS, Hive, and HBase using Sqoop.
* Experienced in processing real-time data using Kafka 0.10.1 producers and stream processors and implemented stream processing using Kinesis and data landed into Data Lake S3.
* Implemented various algorithms for analytics using Cassandra with Spark and Scala.
* Strong Experience in implementing Data warehouse solutions in Confidential Redshift.
* Worked on various projects to migrate data from on-premises databases to Confidential Redshift, RDS, and S3.
* Adept in statistical programming languages like R and Python including Big Data Technologies like Hadoop 2, HAVE, HDFS, MapReduce, and Spark.
* Expertise in building CI/CD on AWS environment using AWS Code Commit, Code Build, Code Deploy, and Code Pipeline and experience in using AWS CloudFormation, API Gateway, and AWS Lambda in automation and securing the infrastructure on AWS.
* Created Automation to create infrastructure for Kafka clusters with different instances as per components in the cluster using Terraform for creating multiple EC2 instances & attaching ephemeral or EBS volumes as per instance type in different availability zones & multiple regions in AWS.
* Experienced in configuring and administering the Hadoop Cluster using major Hadoop Distributions like Apache Hadoop and Cloudera.

**Professional Experience**

**KPMG, Tampa, FL August 2022 - Present**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Experienced in writing Spark Applications in Scala and Python.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators.
* Experience in GCP Dataproc, GCS, Cloud functions, Big Query.
* Analyzed large and critical datasets using HDFS MapReduce, Kafka, Spark, HBase, Hive, Hive UDF, and Spark.
* Used Kafka consumer’s API in Scala for consuming data from Kafka topics.
* Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines
* Developed Python code to gather the data from HBase (Cornerstone) and designed the solution to implement using PySpark.
* Ran data formatting scripts in Java and created terabyte CSV files to be consumed by Hadoop MapReduce jobs.
* Designed, build and managed ELT data pipeline, leveraging Airflow, python, DBT, Stitch Data and GCP solutions.
* Implemented Kafka model which pulls the latest records into Hive external tables.
* Loaded all datasets into Hive from Source CSV files using spark and Cassandra from Source CSV files using Spark/PySpark.
* Developed Cloud Functions in Python to process JSON files from source and load the files to Big Query.
* Involved in implementing security on Hortonworks Hadoop Cluster using Kerberos by working along with operations team to move nonsecure cluster to secured cluster.
* Skilled in leveraging **cl**oud-native services and APIs 2.0 to integrate applications with GCP and AWS, enabling efficient data processing and service orchestration.
* Created Build and Release for multiple projects in a production environment using Visual Studio Team Services. Implemented schema extraction for Parquet and Avro file Formats in Hive/MongoDB.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Written Spark applications using Scala to interact with the PostgreSQL database using Spark SQL Context and accessed Hive tables using Hive Context.
* Used Spark Streaming to receive real-time data from Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra.
* Experience in integrating real-time data using Kafka and Spark Streaming in Java and Scala to process it coming from different source systems such as XLSX, CSV, Database, etc.
* Developed PIG and Hive UDFs in java for extended use of PIG and Hive and wrote Pig Scripts for sorting, joining, filtering, and grouping the data.
* Excellent problem-solving and troubleshooting skills, demonstrated through performance monitoring, optimization, and issue resolution in GCP and AWS environments.
* Worked and troubleshot in establishing the connections between Cognos BI and AWS Athena.
* Automated buildings using Maven and scheduled automated nightly builds using Jenkins. Built Jenkins pipeline to drive all microservices builds to the Docker registry and then deployed to Kubernetes.
* Involved in ETL, Data Integration, and Migration by writing Pig scripts.
* Exported the analyzed data to Teradata using Sqoop for visualization and to generate reports for the BI team. · Migrated the computational code in HQL to PySpark.
* Imported data into HDFS from various SQL databases and files using Sqoop and from streaming systems using Storm into Big Data Lake.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Work related to downloading BigQuery data into pandas or Spark data frames for advanced ETL capabilities · Completed data extraction, aggregation, and analysis in HDFS by using PySpark and storing the data needed to Hive.
* Exposure to the usage of Apache Kafka develops a data pipeline of logs as a stream of messages using producers and consumers.
* Sound knowledge in programming Spark using Scala.
* Populated HDFS and HBase with huge amounts of data using Apache Kafka.
* Developed spark applications in Spark on the distributed environment to load a huge number of CSV files with different schema into Hive ORC tables.
* Designed SSIS Packages to transfer data from flat files, Excel SQL Server using Business Intelligence Development Studio. · Developed a POC for project migration from the on-premises Hadoop MapR system to GCP.
* Analyzed the SQL scripts and designed the solution to implement using Pyspark.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, and Big Query.
* Coordinated with the team and developed a framework to generate Daily Ad hoc reports and Extracts from enterprise data from Big Query.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Experienced in working with various kinds of data sources such as Teradata and Oracle. Successfully loaded files to HDFS from Teradata and load loaded from HDFS to Hive and Impala.
* Installed Oozie workflow engine to run multiple Hive and Pig jobs which run independently with time and data availability.
* Involved in POC Data Extraction, aggregations, and consolidation of data within AWS Glue using PySpark.
* Supported customer service designing ETL jobs dashboards utilizing data from Redshift.
* Responsible for real-time heterogeneous replication from MySQL to Oracle DB and MySQL to SQL Redshift using the Tungsten replicator tool while performance tuning and maintaining both databases.
* Created Redshift clusters on AWS for quick accessibility for reporting needs.
* Designed and deployed a Spark cluster and different Big Data analytic tools including Spark, Kafka streaming, AWS, and HBase with Cloudera Distribution.

**Environment**: Spark RDD. Python, Hadoop, AWS Glue, Apache Kafka, Amazon S3, SQL, Spark, AWS cloud, GCP, Big query, Kafka, SQL, Cloudera, Cassandra, Tableau, NoSQL, redshift, ETL, and Spark (PySpark, MLlib, EMR, EC2, and Amazon RDS), Data lake, Cloudera Stack, HBase, Hive, Impala, Pig, NiFi, Spark, Spark Streaming, Elasticsearch, Teradata, Logstash, Kibana, JAX-RS, Spring, Ubuntu, and Solar.

**Charter Communications, St. Louis, MO July 2020 - July 2022**

**Role: Data Engineer**

**Responsibilities:**

* Integrated data from various system source flat files, CSV, and Oracle using PySpark.
* Created efficient documentation for process/procedure rules.
* Implemented data pipeline for processing data objects into S3 data bucket with security policies and ensuring uniformity of data.
* Developed test scripts using python and PostgreSQL for data validation.
* Written the python scripts to load the data into the tables.
* Used Autosys and Move IT tools to place the file from on premises to AWS-S3 buckets.
* Written the jil scripts to AutoSys jobs and scheduled jobs on every business day.
* Used celery scheduler to places the file into tables from AWS-S3.
* Used AWS Athena to query the data on top s3 bucket (files in inbound) which helps the businesspeople for data analysis. Rewrote existing python/Flask module to deliver certain format of data.
* Created script in python for calling REST APIs.
* Deployed Django web application in Apache webserver and cloud web deployment.
* Installed Kafka manager for consumer lags and for monitoring Kafka metrics also this has been used for adding topics, Partitions etc.
* Proficient in building data pipelines using Apache Airflow in GCP for ETL-related jobs, using different airflow operators.
* Involved in designing user interactive web pages as the front-end part of the web application using various web technologies like HTML, JavaScript, jQuery and implementing CSS for better appearance and feel.
* Skilled in analyzing SQL scripts and designing solutions for implementation using PySpark, with experience configuring GCP services such as Data Proc, Storage, and Big Query using cloud shell SDK.
* Used Celery as task queue and RabbitMQ, Redis as messaging broker to execute asynchronous tasks.
* Worked on MongoDB database concepts such as locking, transactions, indexes, Sharding, replication, schema design.
* Written PostgreSQL sequels according to the business transformations to place the data into respective tables.
* Created the Splunk dashboard for run time errors directed from AWS cloud using kinesis.
* Involved in software development life cycle (SDLC) of tracking the requirements, gathering, analysis, detailed design, development, system testing and user acceptance testing.
* Actively involved in developing the methods for Create, Read, Update and Delete (CRUD) in Active Records.
* Developed entire frontend and backend modules using Python on Django Web Framework by implementing MVC architecture.
* Implemented cloud integrations to GCP and Azure for bi-directional flow setups for data migrations.
* Managed datasets using Panda data frames and MySQL, queried MYSQL database queries from python using python-MySQL connector and MySQL DB package to retrieve information.
* Worked with external vendors or partners to onboard external data into Target S3 buckets.
* Worked on Oozie to develop workflows to automate the ETL data pipeline.
* Expertise in building and architecting end-to-end data pipelines and ETL/ELT processes for data ingestion and transformation in GCP.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Perform Data Cleaning, features scaling and features engineering using pandas and NumPy packages in python.
* Successfully set up a no authentication Kafka listener in parallel with Kerberos (SASL) Listener. Also, I tested a non-authenticated user (an Anonymous user) in parallel with the Kerberos user.
* Advanced knowledge of Confidential Redshift and MPP database concepts.
* Migrated on-premises database structure to Confidential Redshift data warehouse.
* Implemented Workload Management (WML) in Redshift to prioritize basic dashboard queries over more complex longer-running ad hoc queries. This allowed for a more reliable and faster reporting interface, giving sub-second query responses for basic queries.
* Day-to-day responsibility includes developing ETL Pipelines in and out of the data warehouse and developing major regulatory and financial reports using advanced SQL queries in snowflake.
* Stage the API or Kafka Data (in JSON file format) into Snowflake DB by Flattening the same for different functional services.
* Worked on continuous Integration tools Jenkins and automated jar files at the end of the day.
* Heavily involved in testing Snowflake to understand the best possible way to use the cloud resources.
* Worked extensively on AWS Components such as Airflow, Elastic Map Reduce (EMR), Athena, and Snowflake.
* Implemented various Data Modeling techniques for Cassandra.
* Possess good knowledge in creating and launching EC2 instances using AMIs of Linux, Ubuntu, RHEL, and Windows and wrote shell scripts to the bootstrap instances.
* Design an ELK system to monitor and search enterprise alerts. Installed, configured, and managed the ELK Stack for Log management within EC2 / Elastic Load balancer for Elastic Search.
* Loaded data into redshift tables from s3, DynamoDB, EMR, and EC2 using copy utility.
* Worked on MongoDB database concepts such as locking, transactions, indexes, Sharding, replication, and schema design. Created multiple databases with shared collections and chose shard key based on the requirements. Experience in managing MongoDB environment from availability, performance, and scalability perspectives.
* Applied write concern for the level of acknowledgment while MongoDB writes operations to avoid rollback.
* Integrated with UI layer using HTML, Ajax, JavaScript.
* Created multi-tier java-based multiple web services to read data from MongoDB.
* Worked with Lead in NoSQL column family design, client access software, and Cassandra tuning during migration from Oracle-based data stores.
* CAT EDH team's goal is to get rid of all their Teradata stacks and have all data on Amazon’s S3 storage platform.
* Used AWS Redshift, S3, Spectrum, and Athena services to query large amounts of data stored on S3 to create a Virtual Data Lake without having to go through the ETL process.
* Automate AWS infrastructure through infrastructure as code by writing various Terraform modules and scripts by creating AWS IAM users, groups, roles, policies, custom policies, AWS Glue, Crawlers, Redshift clusters, snapshots of clusters, EC2, and S3 buckets.
* Published Power BI Reports in the required originations and Made Power BI Dashboards available in Web clients and mobile apps.
* Create ETL Mappings for the Operational dashboard for various KPIs and Business Metrics, allowing powerful drill down for Detail reports to understand the data at a very detailed level.
* Used Informatica Power Center Workflow manager to create sessions, workflows, and batches to run with the logic embedded in the mapping.

Develop ETL pipelines in and out of the data warehouse using a combination of Python and Snowflake’s Snow SQL

* Extract Transform and Load data from sources Systems to Azure Data Storage services using a combination of Azure Data factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Implemented Security roles on SSAS cubes and Power BI services by implementing authentication and authorization methods.
* Worked with BI Developer, I was involved in developing different reports using Microsoft BI tools like SSIS and SSRS and developed Power BI dashboards that would help businesses operate smoothly.
* Used APIs with Python to extract and push prediction outputs in EMR.
* Created continuous integration and continuous delivery (CI/CD) pipeline on AWS that helps to automate steps in the software delivery process.
* Upgraded from Cloudera manager and CDH from 5.7.1 to CDH 5.11.2, worked with Cloudera to set up SSO login and worked on data migration in HIVE. Worked on Importing and exporting data into HDFS and Hive using Sqoop.

**Environment**: CSV, Oracle, PySpark, Spark RDD, S3. JSON, Parquet, ETL, Hive, UNIX, HDFS, Power BI, SQL Server. DevOps, Microsoft Azure, Kerberos, Scala, Python, AWS, Kafka, Redshift, Airflow, Snowflake, Cassandra, Map Reduce (EMR), AWS Athena, Linux, Ubuntu, EC2, MongoDB, JavaScript, Ajax, CAT, GCP, AWS Glue, Crawlers, Power Center Snow SQL, Cloudera, Jenkins. data bricks.

**Viacom CBS, NYC, NY December 2019 - June 2020**

**Role: Data Engineer**

**Responsibilities:**

* Designed batch processing jobs using Apache Spark to increase speed compared to MapReduce jobs.
* Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation.
* Automated all the jobs for pulling data from the FTP server to load data into Hive tables, using Oozie workflows.
* Involved in scheduling Oozie workflow engine to run multiple HiveQL, Sqoop, and Pig jobs.
* Designed HBase row key and data modeling to insert to HBase tables using concepts of lookup tables and staging tables.
* Developed data pipeline using Sqoop to ingest customer behavioral data and purchase histories into HDFS for analysis.
* Collected the JSON data from HTTP Source and developed Spark APIs that help to do inserts and updates in Hive tables.
* Developed shell scripts for running Hive scripts in Hive and Impala.
* Designed new VPCs for the secure access POCs, and architected and implemented multiple Amazon Glue/Spark pipelines with 100s DPUs.
* Involved in importing the real-time data to Hadoop using Kafka and implemented the Oozie job for daily imports.
* Used Azure Data Factory as an orchestration tool for integrating data from upstream to downstream systems.
* Automated jobs using different triggers (Event, Scheduled, and Tumbling) in ADF.
* Used Cosmos DB for storing catalog data and event sourcing in order processing pipelines.
* Designed and developed user-defined functions, stored procedures, and triggers for Cosmos DB
* Created DA specs and Mapping Data flow and provided the details to the developer along with HLDs.
* Implemented AWS Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using Docker Automation techniques.
* ETL Restarting capability for a date or date range or from the point of failure or from the beginning.
* Extract data from different social analytic sites (Facebook, Google+, Twitter, YouTube, Ooyala, iTunes, Google Analytics, Sony games) through Pentaho DI (ETL)
* Conducted statistical analysis on healthcare data using python and various tools.
* Responsible for building scalable distributed data solutions using Hadoop.
* Develop a big data web application using Agile methodology in Scala as Scala has the capability of combining functional and object-oriented programming.
* Responsible for Designing Logical and Physical data modeling for various data sources on Confidential.
* Designed and Developed ETL jobs to extract data from Salesforce replica and load it in data mart in Redshift.
* Integrated Cassandra as a distributed persistent metadata store to provide metadata resolution for network entities on the network.

**Environment**: Agile, AWS, Spark, Hive, Oozie, MapReduce, FTP, HiveQL, Sqoop, Pig, Amazon S3, HDFS, Cosmos, ETL scripts, MYSQL, Kafka, Cassandra, T-SQL, Spark SQL, U-SQL, Azure Data Factory. Teradata, Cloud. EBS, EC2, ELB, RDS, S3, SNS, S3, Azure Data bricks, informatica, redshift, python, PySpark, scala, MongoDB, java, Jenkins, data bricks.

**Microsoft, Hyderabad, Telangana, India May 2017 - November 2019**

**Role: Data Engineer**

**Responsibilities:**

* Worked on Snowflake environment to remove redundancy and load real time data from various data sources into HDFS using Kafka.
* Developed data warehouse model in Snowflake for over 100 datasets.
* Designing and implementing a fully operational production grade large scale data solution on Snowflake Data Warehouse
* Work with structured/semi-structured data ingestion and processing on AWS using S3, Python. Migrate on-premises big data workloads to AWS.
* Designed the data aggregations on Hive for ETL processing on Amazon EMR to process data as per business requirement.
* Involved in migration of data from existing RDBMS to Hadoop using Sqoop for processing data, evaluate performance of various algorithms/models/strategies based on real-world data sets.
* Implemented Data Validation using MapReduce programs to remove unnecessary records before move data into Hive tables.
* Implemented cloud integrations to GCP and Azure for bi-directional flow setups for data migrations.
* Created Hive tables for loading and analyzing data and developed Hive queries to process data and generate data cubes for visualizing.
* Extracted data from HDFS using Hive, Presto and performed data analysis using Spark with Scala, PySpark and feature selection and created nonparametric models in Spark.
* Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
* Migrated entire customer data from AWS MySQL RDS to GCP Fire store using python script.
* Worked on Enterprise Messaging Bus with Kafka-Tibco connector and published Queues were abstracted using Spark Dstreams and parsed XML, JSON data in Hive.
* Designed and configured Kafka cluster to accommodate heavy throughput of 1 million messages per second. Used Kafka producer 0.6.3 APIs to produce messages.
* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Used Kafka and Kafka brokers, initiated the spark context and processed live streaming information with RDD and Used Kafka to load data into HDFS and NoSQL databases.
* Entire analytical solution has been built on GCP using various GCP services like Fire store (Firebase), Cloud Storage, Big query, Data Flow, Data Transfer Service (DTS), Cloud Run, Cloud Functions, Pub-Sub, Fire store, Cloud SQL, Cloud Build, Airflow (For orchestration), Apache NiFi etc.
* Subscribing the Kafka topic with Kafka consumer client and processing the events in real time using spark.
* Development of Spark structured streaming to read the data from Kafka in real time and batch modes, apply different mode of Change data captures (CDCs) and then load the data into Hive.
* Developed and Configured Kafka brokers to pipeline server logs data into spark streaming
* Used Apache Kafka to aggregate web log data from multiple servers and make them available in Downstream systems for analysis.
* Integrate AWS Kinesis with on premise Kafka cluster.
* Implemented data ingestion and handling clusters in real time processing using Kafka.
* Developed Sqoop and Kafka Jobs to load data from RDBMS, External Systems into HDFS and HIVE.

**Environment:** Azure Data Factory, Azure ML, Azure blobs, Azure SQL, Azure SQL Data Warehouse, Azure Power Shell, C#, SQL Server, SSMS, TFS, Visual Studio, Visual Studio Online.

**WebSoc Technologies, Hyderabad, Telangana, India July 2015 - April 2017**

**Role: Python Developer/Data Engineer**

**Responsibilities:**

* Familiarity with ORMs such as SQL Alchemy for efficient database interaction.
* Proficient in using Python libraries like Pandas, NumPy, and Matplotlib for data analysis and visualization.
* Experience with Python testing frameworks like unit test, pytest, and others.
* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks, etc.
* Created several Databricks Spark jobs with PySpark to perform several tables-to-table operations. Implemented data pipeline for processing data objects into S3 data bucket with security policies and ensuring uniformity of data. Developed test scripts using python and PostgreSQL for data validation.
* Written the python scripts to load the data into the tables.
* Used Autosys and Move IT tools to place the file from on premises to AWS-S3 buckets.
* Written the jil scripts to AutoSys jobs and scheduled jobs on every business day.
* Used celery scheduler to places the file into tables from AWS-S3. Used AWS Athena to query the data on top s3 bucket (files in inbound) which helps the businesspeople for data analysis.
* Importing and exporting data jobs to perform operations like copying data from HDFS and to HDFS using Sqoop and developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Use Kafka a publish-subscribe messaging system by creating topics using consumers and producers to ingest data into the application for Spark to process the data and create Kafka topics for application and system logs.
* Advanced knowledge of Confidential Redshift and MPP database concepts.
* Analyze escalated incidences within the Azure SQL database. Implemented test scripts to support test-driven development and continuous integration.

**Environment**: SQL, Python, NumPy, Unit Test, Azure, Data Bricks, PySpark, PostgreSQL, AWS, HDFS, Sqoop, Kafka, Redshift, MPP, TDD.

**Arissoft Solutions, Hyderabad, India June 2014 - June 2015**

**Role: Python Developer**

**Responsibilities:**

* Skilled experience in Python with proven expertise in using new tools and technical developments (libraries used: libraries- psutil, Kafka python, NumPy, SciPy, matplotlib, Pickle, python-twitter, Panda’s data frame, networks, urllib2, MySQL dB for database connectivity)
* Experienced in working on Application Servers like WebLogic, Tomcat, Web Servers like Apache server, NGINX, and Integrated Development Environments like PyCharm, PyDev, and Eclipse.
* Developed tools using Python, Shell scripting, XML to automate some of the menial tasks. Interfacing with supervisors, artists, systems administrators, and production to meet production deadlines.
* Write scripts using python modules and their libraries to develop programs that improve the processing of access requests.
* Worked on PowerShell for automation and made use of command-line shell and shell scripting.
* Developed Merge jobs in Python to extract and load data into PostgreSQL database, also worked on Python ETL file loading and use of the regular expression.
* Used various SSIS tasks like Bulk Insert Task, Task, Script Task, Send Mail Task.
* Worked on the creation of views to restrict access to data in a table for security and was responsible for writing user-defined functions for mathematical and other operations.
* Designed and developed stored procedure queries necessary to support SSRS reports.

**Environment:** Kafka, Python, NumPy, SciPy, MySQL, WebLogic, Tomcat, Webservers, Apache, Eclipse, XML, Shell Scripting, PostgreSQL, Python ETL, SSIS, SQL, SSRS.

**Education:**

* Bachelor of Technology in Electronics and Communication Engineering (2009-2013).