**Supraj**

**Big Data Engineer**

Mobile: 469 846 8345

Email: suprajsbss@gmail.com

**PROFESSIONAL SUMMARY:**

* Having 11 years of experience in IT which includes Analysis, Design, Development of Big Data using Hadoop, AWS, Scala, design and development of web applications using JAVA, Spring boot and data base and data warehousing development using My SQL, Oracle.
* Experience in transforming business requirements into machine learning analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data and expertise working on variety of industries including healthcare and banking.
* Around 4+ years of work experience on Big Data Analytics with hands on experience in installing, configuring and using ecosystem components like Hadoop Map reduce, HDFS, HBase, Zookeeper, Hive, Sqoop, Pig, Flume, Cassandra, Kafka and Spark, Oozie, Airflow, NiFi(ETL).
* Good understanding of Spark architecture with Databricks, structured streaming. Setting up AWS and Microsoft azure with databricks, databricks workplace with business analytics, manage clusters in databricks.
* Extract, transform and load data from sources systems to azure data storage services using a combination of azure data factory, spark SQL and azure data lake analytics. Data ingestion into one or more azure services like azure data lake, azure storage, azure SQL, azure DW and processing the data in Azure databricks.
* Experience in building and architecting multiple data pipelines, end to end ETL and ELT process for data ingestion and transformation in Google GCP and coordinate task among the business teams.
* Building a Scala and Spark based configure framework to connect data sources like MySQL, Oracle, SQL server, Salesforce and load it in BigQuery.
* Worked on Apache Flink to implement the transformations on data streams for filtering, aggregating and during updating stage.
* Working experience with Apache Flink distributed processing engine for scalable data analytics framework. Used Flink to process large scala data streams which helped in providing data analytical reports for the end clients.
* Working knowledge of Kafka for real time data streaming and event based architecture.
* Experience in Snowflake multi-cluster warehouses.
* Understanding in Snowflake Cloud technology.
* Experience in Snowflake cloud data warehouse and AWS S3 bucket for integrating from multiple source system which include loading JSON formatted data into snowflake table.
* Good Understanding of Hadoop architecture and Hands on experience with Hadoop components such as Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce concepts and HDFS Framework.
* Expertise in preparing Interactive Data Visualization's using Tableau Software from different sources.
* Strong experience in analyzing large amounts of data sets writing PySpark scripts and Hive queries.
* Running of Apache Hadoop, CDH and Map-R distros, dubbed Elastic MapReduce(EMR) on (EC2)
* Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.
* Experienced with cloud: Hadoop-on-Azure, AWS/EMR, Cloudera Manager (also direct-Hadoop-EC2 (non EMR)).
* Extensively worked on AWS services like EC2, S3, EMR, FSx, Lambda, Cloud watch, RDS, Auto scaling, Cloud Formation, SQS, ECS, EFS, DynamoDB, Route53, Glue etc.
* Experience in AWS Eventbridge while using various data events to connect various application components, which was used in building scalable event driven applications.
* Hands on experience in VPN Putty and WinSCP, CI/CD(Jenkins).
* Experienced in writing Map Reduce programs & UDF's for both Pig & Hive in java.
* Experience in dealing with log files to extract data and to copy into HDFS using flume.
* Experience in integrating Hive and Hbase for effective operations.
* Expertise in Waterfall and Agile - SCRUM methodologies.
* Implemented various J2EE standards and MVC framework involving the usage of Struts, JSP, AJAX and servlets for UI design.
* Experienced with code versioning and dependency management systems such as Git, SVT, and Maven.
* Experience in Testing and documenting software for client applications.
* Writing code to create single-threaded, multi-threaded or user interface event driven applications, either stand-alone and those which access servers or services.
* Good experience in using Data Modelling techniques to find the results based on SQL and PL/SQL queries.
* Experience working with different databases, such as Oracle, SQL Server, MySQL and writing stored procedures, functions, joins, and triggers for different Data Models.
* Experienced in handling different file formats like Text file, Avro data files, Sequence files, Xml and Json files.

**TECHNICAL SKILLS**:

|  |  |
| --- | --- |
| Big Data Technologies | HDFS, Hive, MapReduce, Pig, Sqoop, Flume, Oozie, Hadoop distribution, and Hbase, Spark, Spark Streaming, Yarn, Zookeeper, Kafka, ETL.(Nifi, Talend etc.) |
| Programming languages | Core Java, Spring Boot, R, Scala, Terraform, Angular. |
| Databases | MySQL, MS-SQL Server 20012/16, Oracle 10g/11g/12c |
| Scripting/Web Languages | HTML5, CSS3, XML, SQL, Shell/Unix, Perl, Python. |
| Databases | Cassandra, HBASE, mongoDB, Oracle, MS SQL, Teradata. |
| Operating Systems | Linux, Windows XP/7/8/10, Mac. |
| Software Life Cycle | SDLC, Waterfall and Agile models. |
| Utilities/Tools | Eclipse, Tomcat, NetBeans, JUnit, SQL, SVN, Log4j, SOAP UI, ANT, Maven, Alteryx, Visio, Jenkins, Jira, Intellij. |
| Data Visualization Tolls | Tableau, SSRS, Cloud Health. |
| Cloud Services | AWS(EC2, S3, EMR, RDS, Lambda, Cloudwatch, FSx, Auto scaling, Redshift,Cloud Formation, Glueetc.) , Azure Databricks, GCP and BigQuery |

**PROFESSIONAL EXPERIENCE:**

**Goldman Sachs , Dallas, TX July 2022 – Till Date**

**Big Data Engineer (Tech Risk, Measurement Engineering)**

* Ingested transactional and various forms of data from primary data sources using Spark, Redshift, S3 and python.
* Familiar with key data science concepts like statistics, data visualization, machine learning etc.
* Experience in Python, R, SAS, PySpark programming for statistic and quantitative analysis.
* Working experience in extracting, transforming data using azure services ranging from azure databricks, azure relational and non relational databases and azure datafactory and clound services.
* Experinecein implementing data architecture using apache spark/flink in production on azure databaricks and AWS sagemaker.
* Responsible for designing and implementing very large scala data engineering solutions using databricks to prepare business reports and working experience with databricks for cloud based data processing and also processing data from various data sources.
* Provisioning Amazon DynamoDB for a fast and flexible NoSQL database service for the applications the need consistent data latency for various tpyes of data.
* Working experience in AWS Eventbridge, used in creating and updating event driven data architectures which helped in creating business reports for end users.
* Developed data ingestion modules using AWS Step functions, AWS Glue and Python modules.
* Worked on both batch time and real time data processing into various layers in S3, Redshift and Snowflake using AWS kinesis, AWS Glue, AWS Lambda and AWS step functions.
* Created data sharing between two snowflake accounts, redesigned the views in Snowflake to increase the performance.
* Using g-cloud function with python to load data in BigQuery for an arrival csv files in GCS bucket.
* Experience in processing and loading data from various data sources into BigQuery using Spark, scala and python and also loading bound and unbound data from google to BigQuery using cloud dataflow using python.
* Design and develop spark job with scala to implement end to end data pipeline for batch data processing.
* Experience in developing enterprise level software and business solutions using batch processing using Apache framework and streaming framework using apache streaming, apache kafka and apache flink.
* Involved in the development of real time streaming applications using Pyspark, Apache flink, Kafka, Hive on distributed Hadoop cluster.
* Performed in Unit testing the data between Redshift and Snowflake and developed data warehouse model in snowflake.
* Created reports in looker based on Snowflake connections and consultaing on Snowflake Data Platform solution architecture, design, development and deployment focused to bring data driven culture across the enterprises.
* Knowledge of Snowflake database, schema and table structures.
* Worked on Measurement Canvas in developing metrics for Vulnerability management for IA metrics mapping.
* Have good knowledge on Metric and measure development and worked on executing scala and spark queries in Zeppelin notebook.
* Worked on Apache Flink to implement the transformations on data streams for filtering, aggregating and during updating stage.
* Install Kafka on Hadoop cluster and configure producer and consumer coding part in java to establish connection from source to HDFS.
* Created data extraction, cleansing and loading to move data from source systems to data warehouses.
* Have experience in updating barrier conditions of the metrics and then successfully running the metrics according to business needs of the organization.
* Worked on Query Analysis of the Adoption metrics to improve the efficiency of the metrics which impacts the end user business needs.
* Collabarated and contributed to identifying project risks, designing mitigation plans by developing and working on GIT upgrade project and streamlining the code using functions and along with using credentials vaults.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Analysed the SQL scripts and designed the solution to implement using Scala.
* Importing and exporting data into HDFS Sqoop and Flume and Kafka.
* Worked with BI team in the area of Big Data Hadoop cluster implementation and data integration in developing large-scale system software.
* Written Hive queries for data analysis to meet the business requirements.
* Used Qlikview and D3 for visualization of query required by BI team.
* Create Hive external tables on the MapReduce output before partitioning, bucketing is applied on it.
* Configured Hive Server (HS2) to enable analytical tools like Tableau, Qlikview and SAS to interact with Hive tables**.**

**Verizon Communications, Plano, Texas                                                            Aug 2021 – July 2022**

**Big data Engineer**

**Responsibilities:**

* Import data from sources like HDFS/HBase into Spark RDD.
* Usage of Spark Streaming and Spark SQL API to process the files.
* Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems/mainframe and vice-versa loading data into HDFS.
* Stored data in AWS S3 like HDFS and performed EMR programs on data stored in S3.
* Worked on Big Data Hadoop cluster implementation and data integration in developing large-scale system software.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Hive.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Involved in Migrating the platform from Cloudera to EMR platform..
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run MapReduce jobs in the backend.
* Extensively involved in developing Restful API using JSON library of Play framework.

**CVS Health Inc, Dallas, TX                                                                   July 2016 – July 2021**

**Lead Big data /AWS Engineer**

* Involved in Requirement gathering, Business Analysis and translated business requirements into Technical design in Hadoop and Big Data
* Involved in SQOOP implementation which helps in loading data from various RDBMS sources to Hadoop systems.
* Worked on Apache Flink to implement the transformations on data streams for filtering, aggregating and during updating stage.
* Working experience with Apache Flink distributed processing engine for scalable data analytics framework. Used Flink to process large scala data streams which helped in providing data analytical reports for the end clients.
* Developed Python scripts to extract the data from the web server output files to load into HDFS.
* Involved in HBASE setup and storing data into HBASE, which will be used for further analysis.
* Created various data pipelines usingSpark, Scala and SparkSQL for faster processing of data.
* Written Spark-SQL and embedded the SQL in SCALA files to generate jar files for submission onto the Hadoop cluster
* Created Data Quality Scripts using SQL and Hive to validate successful das ta load and quality of the data. Created various types of data visualizations using Python and Tableau.
* Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format Parsed Semi Structured JSON data and converted to Parquet using Data Frames in PySpark.
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions, Documented the systems processes and procedures for future references.
* Worked with a team to migrate from Legacy/On prem environment into AWS.
* Created Dockerized backend cloud applications with exposed Application Program Interface (API) interfaces and deployed on Kubernetes.
* Stored and retrieved data from data-warehouses using Amazon Redshift.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Automated and monitored complete AWS infrastructure with terraform.
* Created data partitions on large data sets in S3 and DDL on partitioned data.
* Converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size.

**Sagarsoft, Hyderabad INDIA                                                   Sep 2011 to Oct 2014**

**Software Developer**

**Responsibilities:**

* Excellent JAVA, J2EE application development skills with strong experience in Object Oriented Analysis, extensively involved throughout Software Development Life Cycle (SDLC)
* Implemented various J2EE standards and MVC framework involving the usage of Struts, JSP, AJAX and servlets for UI design.
* Used SOAP/ REST for the data exchange between the backend and user interface.
* Utilized Java and MySQL from day to day to debug and fix issues with client processes.
* Developed, tested, and implemented financial-services application to bring multiple clients into standard database format.
* Assisted in designing, building, and maintaining database to analyze life cycle of checking and debit transactions.

**EDUCATION:**

* Masters in Business Administration (PMBA) from Campbellsville University, 2022.
* Master of Science in Technology Management at Texas A&M University, 2016
* Bachelors from Osmania University, Hyderabad, India, 2011