**NAME:Pruthvi G**

[EMAIL](mailto:EMAIL): saisri@mastroservices.com

PHONE#4698468345

**Summary:**

* Overall 9+ years of IT experience in Analysis, design, development, implementation, maintenance and support with experience in developing strategic methods for deploying big data technologies to efficiently solve Big Data processing requirement
* Experience on BIG DATA using HADOOP framework and related technologies such as HDFS, MapReduce, Spark, HIVE, OOZIE, SQOOP
* Experience in data analysis using HIVE and custom Map Reduce programs
* Expert level experience in designing, building and managing applications to process large amounts of data in a Hadoop ecosystem
* Extensive experience with performance tuning applications on Hadoop and configuring Hadoop systems to maximize performance
* Explored the Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark-Context, Spark-SQL, Data Frame
* Performed analysis on Datamart using spark-SQL joins and window functions.
* Hands of Experience in GCP,BigQuery,GCS bucket, DataProc
* Responsible for implementing a generic framework to handle different data collection methodologies from the client primary data sources, validate transform using spark and load into S3 or GCP buckets
* Experience on SPARK, SCALA in GCP environment
* Developed analytical components using SCALA, SPARK, HBASE and SPARK STREAM
* Performed Hadoop backup Strategy to take the backup of hive, HDFS, HBase, oozie, Airflow etc.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems (RDBMS) and from RDBMS to HDFS.
* Created the HBASE tables for validation, audit and offset management table.
* Created logical view instead of tables in order to enhance the performance of HIVE queries
* Involved in developing Hive DDLS to create, alter and drop Hive tables
* Pretty Good Knowledge on HIVE Optimization techniques like Vectorization and column-based optimization
* Written OOZIE, AIRFLOW workflow to invoke the Jobs in predefined Interval
* On Other Hand working on POC with Kafka and NIFI to pull the real-time events into Hadoop Box
* Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, SPARK-SQL, DATA FRAME, PAIR RDD’s and YARN
* Developed Spark Structured Streaming Job and created STREAMING DATA FRAME
* Experience in creating BATCH DATA FRAME and STREAMING DATA FRAME
* Experience with migration to Amazon web Services (AWS) from HDFS
* Implemented security by using IAM &amp; private key in AWS cloud
* Solid understanding of RDBMS database concepts including performance tuning and query optimization
* Experience in Java, J2EE, Java Scripting, HTML, JSP
* Solid programming knowledge on SCALA
* Good working experience on Hadoop tools related to Data warehousing like Hive and also involved in extracting the data from these tools on to the cluster using Sqoop.

**Technical Skills:**

* **Hadoop/Spark Ecosystem:**Hadoop, Map Reduce, Hive, YARN, Flume, Sqoop, Oozie, Zookeeper, Spark, Airflow, Cassandra, HBase.
* **Hadoop Distribution:**Cloudera distribution and Horton works.
* **Cloud: GCP –** Big Query, GCP buckets, dataProc clusters AWS – S3, EC2, EMR
* **Programming Languages:**Scala, Java, SQL, HQL
* **Databases:**Oracle, MySQL, SQL Server, MS Access, HBase, Cassandra, MongoDB
* **Operating Systems:**Linux, Windows, Ubuntu.
* **IDE:**IntelliJ, Eclipse,Pycharm
* **Programming Language:** C, C++, Java, Python.
* **Scripting Language:** JSP, HTML, XML.
* **Version Control:** BitBucket,git
* **Methodology:** Agile, Waterfall.

**Professional Experience:**

**Client : Mass Mutual,springfield,MA**

**Designation : Big Data Developer/GCP**

**Duration : Apr 2022 – Till date**

**Roles & Responsibilities:**

* Hands of Experience in GCP,BigQuery,GCS bucket, DataProc.
* Designed, implemented and developed ETL solutions for data ingestion, cleansing, business rules execution as perthe business requirements.
* Big data Developer and developed code with various transformations using hive sql and spark scala .
* Expertise On optimizing spark Jobs in GCP.
* Written oozie workflow to invoke the Jobs in predefined Intervals.
* Expertise in Hive/SQL queries into Spark transformations using Spark RDDs and Scala
* Experienced in working with spark eco system using Spark SQL and Scala queries on different formats like Textfile, CSV file and Performed necessary Transformations and Aggregation on data model and persists the data inHDFS.
* Familiar with hashing the sensitive data using Jenkins job.
* Knowledge and experience to use code versioning tools such as Git.
* Experience with databases, PostgreSQL preferred and Oracle,
* Team player and individual contributor.
* Ability to accurately identify root cause of technical problems.
* Used Hive to analyze the Partitioned and Bucketed data and compute various metrics for reporting.
* Expertise On optimizing spark Jobs when dealing with Huge joins and data Skew.
* Written oozie workflow to invoke the Jobs in predefined Intervals.
* Performed necessary Transformations and Aggregation on the fly to build the common learner data model andpersists the data in HDFS.
* Explored the usage of Spark for improving the performance and optimization of the existing algorithms in Hadoopusing Spark Context, Spark SQL and Spark Yarn.
* Used Hive to analyze the Partitioned and Bucketed data and compute various metrics for reporting.
* Involved in converting Hive/SQL queries into Spark transformations using Spark dataframes and Scala
* **Production Support:**
* Responsible for monitoring the production servers, scheduled jobs, incident management and receiving incidentsand requests from end-users.
* Analyze the available data and find the root cause of the problem.
* Analyzing the requests and either responding to the end user with a solution or escalating it to the other IT teams.
* Ability to prioritize work to successfully deliver service to agreed levels in a diverse and constantly changingtechnical environment.

**Environment:**  Spark, Hive, Spark SQL, Oozie, Scala, Maven, Jupiter Notebook, GCP, GCS buckets, Bigquery, Dataproc Clusters, Unix Shell Scripting, PostgresSQL, Intellij.

**Client : Trimble, TX**

**Designation : Bigdata AWS Cloud Engineer**

**Duration : Oct 2018 to Apr 2022**

**Roles & 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 RelationalDatabase 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-scalesystem software
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python andScala
* Developing UDFs in java for hive and pig and worked on reading multiple data formats on HDFSusing Scala.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processingwith 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.
* Developed analytical component using Scala, Spark and Spark Streaming.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFSfor further analysis
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke andrun MapReduce jobs in the backend.
* Extensively involved in developing Restful API using JSON library of Play framework.
* Developed Storm topology to ingest data from various source into Hadoop Data Lake.
* Developed web application using HBase and Hive API to compare schema between HBase and Hivetables.
* Played a vital role in Scala/Akka framework for web based applications
* Connected to AWS S3 using SSH and ran spark-submit jobs
* Developed Python Script to import data SQL Server into HDFS &amp; created Hive views on data in HDFSusing Spark.
* Expert in Troubleshooting MapReduce Jobs.
* Troubleshooting data load issues.
* Created scripts to append data from temporary HBase table to target HBase table in Spark.
* Developed complex and Multi-step data pipeline using Spark.
* Worked on Big Data Integration and Analytics based on Hadoop, SOLR, Spark, Kafka, Storm and webMethods technologies.
* Populated HDFS and Cassandra with huge amounts of data using Apache Kafka.
* Monitoring YARN applications. Troubleshoot and resolve cluster related system problems.
* Upgrading the Hadoop Cluster from CDH3 to CDH4, setting up High Availability Cluster andintegrating HIVE with existing applications.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFSfor further analysis.
* Involved in creating ETL flow using Pig, loading with data and writing Pig Latin queries which will runinternally in Map Reduce way.
* Involved in writing Unix/Linux Shell Scripting for scheduling jobs and for writing pig scripts and hiveQL.
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke andrun MapReduce jobs in the backend.
* Assisted in exporting data into Cassandra and writing column families to provide fast listing outputs.
* Used Zookeeper for providing coordinating services to the cluster.
* Worked with Hue UI in scheduling jobs with ease and File browsing, Job browsing, Metastoremanagement.
* Developed and designed system to collect data from multiple portal using kafka and then process itusing spark.

**Environment:** Hadoop, HDFS, Hive, Core Java, Sqoop, NIFI, Spark, Scala, Hive, Cloudera CDH4, Oracle, Elastic search, Kerberos, Datastage, SFTP, data Lake, Impala, Jira, Wiki, Alteryx, Teradata, Shell/Perl Scripting, Kafka, AWS EC2, S3, EMR, Cloudera.

**Client : AIG, TX**

**Designation : Big Data Developer**

**Duration : Sep 2016 to Oct 2018**

**Roles & Responsibilities:**

* Importing and exporting data into HDFS and Hive using Sqoop.
* Used Bash Shell Scripting, Sqoop, AVRO, Hive, Pig, Java, Map/Reduce daily to develop ETL, batchprocessing, and data storage functionality
* Migrated their Big Data Platform from on-premise Hadoop to Google Cloud Platform (GCP) to one ofthe projects that we are working
* Evaluated and designed the software and the infrastructure for the company-wide private containercloud solution based on Kubernetes, Swarm, Docker, Python and Go.
* Managing Linux Containers using Docker and docker registry
* Used Pig to do data transformations, event join sand some pre-aggregations before storing the dataon the HDFS.
* Developed an enterprise application, Spring/Hibernate ORM framework for backend, Spring MVC formiddle layer.
* Worked on Amazon Web Service(AWS) to integrate EMR with Spark 2 and S3 storage and Snowflake
* Open SSH tunnel to GCP-Google DataProc to access to yarn manager to monitor spark jobs.
* Configured Spark streaming to receive real time data from the Kafka and store the stream data intoAWS S3 using Scala.
* Extensively used various Spring modules – AOP (logging/messaging), DI, Auto wiring, Inheritance,JDBC templates, used DAO pattern for fetching data from databases through Hibernate and carrythrough business logic.
* Exploited Hadoop MySQL-Connector to store Map Reduce results in RDBMS.
* Analyzed large amounts of data sets to determine optimal way to aggregate and report on it.
* Worked on loading all tables from the reference source database schema through Sqoop.
* Worked on designed, coded and configured server side J2EE components like JSP, AWS and JAVA.
* Collected data from different databases (i.e. Oracle, MySQL) to Hadoop
* Used Oozie and Zookeeper for workflow scheduling and monitoring.
* Worked on Designing and Developing ETL Workflows using Java for processing data in HDFS/HBaseusing Oozie.
* Experienced in managing and reviewing Hadoop log files.
* Involved in loading and transforming large sets of structured, semi structured and unstructured datafrom relational databases into HDFS using Sqoop imports.
* Working on extracting files from MySQL through Sqoop and placed in HDFS and processed.
* Supported Map Reduce Programs those running on the cluster.
* Cluster coordination services through Zookeeper.
* Involved in loading data from UNIX file system to HDFS.
* Created several Hive tables, loaded with data and wrote Hive Queries in order to run internally inMapReduce.
* Developed Simple to complex MapReduce Jobs using Hive and Pig.

**Environment:** AWS, Big Data Warehouse,Spark, Java, Hive, Stream-sets Spark SQL, Kafka, EMR, Snowflake, Nebula, Hive, Python, GCP, Scala, kubernetes, docker, Maven, Jupiter Notebook, Snowflake, Teradata, Visual Studio, Unix Shell Scripting.

**Client : Cybage Software Private Limited, Hyd, India**

**Designation : Data Engineer**

**Duration : Nov 14 to Dec 15**

**Roles & Responsibilities:**

* Spearheaded analysis, design and development for building a common architecture and data repositoryfor Apparel and Footwear sourcing data across the Geo&#39;s for data management and reporting needs.
* Evaluated, extracted and transformed data for analytic purpose within the context of Big dataenvironment and structured large data sets by applying standard data modelling methods.
* Designed and developed efficient PySpark programs using cloud-based data platforms (EMR) toextract/transform/load data in between various data warehouse applications.
* Worked on Cloud Platform like AWS and possess good knowledge on different types of instances foroptimal usage of clusters based on the requirement.
* Engineered a solution to optimize the ETL process of Alteryx to Snowflake ingestion process whichtakes around 7+ hours into a more reliable process by using PySpark which simplified and drasticallyreduced the data ingestion time to 7 minutes.
* Designed Hive tables over the parquet and csv files for loading and analyzing data.
* Involved in developing DAGS using Airflow orchestration tool and monitored the weekly processes.
* Identify gaps in data processes and drive improvements via continuous improvement loop by ensuringgood data flow between databases and backend systems.

**Environment:** Python, Alteryx, AWS S3, AWS EMR, Spark, Hive, Airflow, Sqoop, Teradata SQL assistant, Snowflake, MS SQL Server 2016, Tableau 2018.3, Snowflake database, R - Programming, SharePoint, SSIS