**Sudharshan**

**Phone:** (978) 315-0436

**E-mail:** sudharshan2183@gmail.com

**LinkedIn:** [**http://www.linkedin.com/in/sudharshan-s-a01371293**](http://www.linkedin.com/in/sudharshan-s-a01371293) **SENIOR DATA ENGINEER**

**Professional Summary**

* Over 8 years of extensive hands-on experience in IT industry, Spark, Scala, Python, Machine Learning Algorithms Deployment, AWS, Apache Nifi, Kafka and Hadoop Components.
* Experience in Building End to end Pipelines for Real time data analytics on Cloud using AWS services EMR, EC2, Dynamo DB , RDS, Athena , S3, Lambda, SNS, SQS.
* Extensive experience in project life cycle including Data Acquisition, Data Cleaning, Validation, Data Manipulation, Data Validation, Data Mining, Algorithms, and Visualization.
* Experience in building Machine Learning models and Wide range exposure to Python Libraries.
* Good experience working with Python oriented to data manipulation, data wrangling and data analysis using libraries like Pandas, NumPy, Scikit-Learn and Matplotlib.
* Experience in Spark, Data Frames, PySpark, Pandas, Spark Streaming, Spark MLIB.
* Experienced Good understanding of NoSQL databases and hands on work experience in writing applications No SQL Databases HBase, Cassandra and MongoDB.
* Monitored cluster for performance and, networking and data integrity issue.
* In-depth knowledge of XML and XSLT data and processing tools like spark-xml of Databricks.
* Experience working with Sequence files, ORC, AVRO file, Parquet file and XML formats.
* Experienced with apply XSL tranformations and Unmarshelling xml documents.
* Hands on experience designing and building data models and data pipelines on Data Warehouse focus and

Data Lakes.

* Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft

Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters In Databricks, Managing the

Machine Learning Lifecycle.

* Supporting Continuous storage in AWS using Elastic Block Storage, S3, Glacier. Created Volumes and configured Snapshots for EC2 instances
* Implementations of generalized solution model using AWS SageMaker.
* Good Hands-on Experience on NoSQL databases like MongoDB, Cassandra, and HBase.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), Data Factory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB),SQL Server, Oracle,Data Warehouse etc. Build multiple Data Lakes.
* Experience creating Visual report, Graphical analysis and Dashboard reports using Tableau, Informatica of historical data saved in Hdfs and data analysis using Splunk enterprise edition.
* Experience in creating, debugging, scheduling and monitoring jobs using Airflow.
* Have good experience creating real time data streaming solutions using Spark Streaming and Kafka.
* Experience working with Snowflake for running data pipelines that has huge volumes.
* Working experience with version control tools like SVN, Git, GitHub and BitBucket.

**Technical Skills**

|  |  |
| --- | --- |
| **Big Data Technologies** | HDFS, Hive, AWS, MapReduce, Pig, Sqoop, Hbase, Airflow, Zookeeper, Yarn, Avro, Spark, Apache Kafka, SQS, Nifi |
| **Databases** | Snowflake, Teradata, Oracle, MySQL |
| **NoSQL Databases** | HBase, MongoDB, Cassandra |
| **Programming languages** | SQL, PL/SQL, HQL, Python, Pyspark, Java and UNIX shell, Scala |
| **Cloud** | AWS S3, AWS EC2, AWS EMR, AWS Airflow, SQS, AWS RDS and AWS Glue. Athena, Lambda, Cloud Watch, Azure, Azure Databricks, Azure Data Explorer, Azure HDInsight |
| **Methodologies** | Waterfall, Agile |
| **Defect Management** | Jira, Quality Center |
| **Operating systems** | Linux, UNIX, MAC, Windows |
| **Web/Application Servers** | Apache Tomcat, Web Logic, Web Sphere |
| **Development Tools** | Pycharm, TOAD, SQL Developer, Ms Office, Eclipse, VM ware, JIRA, CVS, SVN, GIT, Bitbucket, Soap UI, Hue, Dreamweaver, Putty, Winscp |

**Professional Experience**

**Client: Broadridge - Lake Success, NY July 2021 to Present**

**Sr.Data Engineer**

**Responsibilities:**

* Involved in design, develop and support data pipelines using Sqoop, Pig, Spark and Hive and store the data on HDFS and S3.
* Streaming Xml messages through Spark Stremaing, converting Xml formats to various file formats.
* Strong knowledge on XML Schema, GJXML, XSLST.
* Optimized Large Scale Spark Batch jobs that run on Tera bytes of data, one from 26 hours 3.5 hours, 13 hours to 45 minutes by repartitioning Skewed data, by applying different optimization techniques.
* Hands - on experience in Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake.
* Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
* Have good experience working with Azure BLOB and Data lake storage and loading data into Azure SQL Synapse analytics (DW).
* Designed the ETL process and created the high-level design document including the logical data flows, source data extraction process, the database staging and the extract creation.
* Created custom kafka connect source connecter to get data from JMS queues using kafka-connect api and deployed kafka-streams job for data transformation.
* Used Spark MLIB to predict Customer Demand for certain products for long-weekend sales and created score for high demand products which helped to manage store inventory to accommodate the Customer's demand.
* Used Pyspark for data ingestion and perform complex transformations.
* Worked on importing data from MYSQL database to HDFS and vice-versa using SQOOP.
* Responsible for developing Kafka Producers and Consumers from scratch as per the requirement specifications.
* Developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Highly skilled in integrating Kafka with Spark streaming for high-speed data processing.
* Used Spark Dataframes, Spark-SQL extensively to build multiple ETL pipelines.
* Converted RDD's to data frames to improve the performance and optimization using in-memory procedures with Spark Context, Spark-SQL, Data Frame, and Pair RDD's.
* Performance tuning using Partitioning and Bucketing of Hive tables.
* Optimized Java micro services and integrated Data sources Hive and Hbase, reduced latency by 27%
* Load and transform large sets of structured, semi structured and unstructured data that we receive from various vendors.
* Involved in moving data from SFTP server to HDFS and S3 for data processing using Hive and Spark
* Used Airflow for scheduling and orchestrating the data pipelines.
* Conducted requirements gathering sessions with various stakeholders.
* Hands on Experience working with AWS stack such as EMR, EC2, S3, RDS, Lambda for building fault tolerant applications.
* Worked with Tableau developers in optimizing their queries to develop the dash boards

.

**Environment*:*** *Hadoop, AWS, EC2, EMR, S3, HDFS, MapReduce, Spark, Pig, Hive, Impala, Sqoop, Kafka, HBase, Airflow, Tableau, Python, PL/SQL, Snowflake, Teradata, Linux shell scripting, Pyspark, Pycharm, Soap UI, Eclipse, Jenkins, Jira*

**Client: DTCC - Jersey City, NJ Sep 2018 to July 2021**

**Sr.Data Engineer**

**Responsibilities:**

* Hands on experience in installation, configuration, supporting and managing Hadoop Clusters.
* Knowledge of Cassandra security, maintenance and tuning both database and server.
* Chipped away at outlining and building up the Real Time Analysis module for Analytic Dashboard utilizing Cassandra, Kafka, Spark Streaming.
* Installed and configured Confluent Kafka in R&D line. Validated the installation with HDFS connector and Hive connectors.
* 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 Big Data Analytics and Advanced Data Science techniques to identify trends, patterns, and discrepancies on petabytes of data by using Azure Databricks, Hive, Hadoop, Python, PySpark, Spark SQL, MapReduce, and Azure Machine Learning.
* Set-up configured and optimized the Cassandra cluster. Developed real-time Spark based application to work along with the Cassandra database.
* Integrated Kafka with Spark Streaming to listen onto multiple Kafka Brokers with different Kafka topics for every 5 Seconds.
* Extracting defined values from Raw Xml mssages, Used extensive XML Libraries in Java.
* Enhanced and optimized product Spark code to aggregate, group and run data mining tasks using the Spark framework and handled Json Data.
* Handled Json Data comes from Kafka Direct Stream on each partitions and transformed them into required Data Frame Formats.
* Upgraded Spark 1.6 to latest Version Spark 2.2 and configure
* Worked on Import & Export of data using ETL tool Sqoop from MySQL to HDFS.
* Worked on Lambda Architecture for both Batch processing and Real Streaming purposes.
* Appended the Data Frames into Cassandra Key Space Tables using DataStax Spark-Cassandra Connector.
* Configured Authentication and security in Apache kafka pub-sub system.
* Implement and test integration of BI (Business Intelligence) tools with Hadoop stack.
* Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, HBase, Zookeeper, Sqoop, Yarn, Spark2, Kafka and Oozie.
* Formulated procedures for installation of Hadoop, Spark2 patches, updates and version upgrades.

**Environment:** *Cloudera, HDFS, Spark, Hive, Pig, Map Reduce, Hue, Sqoop, Putt, Apache Kafka, Apache Drill,*

*Century Link Cloud, AWS, Java Netezza, Cassandra, Oozie, Spark, SPARK SQL, Maven, SBT, Java, Scala, SQL and Linux, YARN, Agile Methodology, Solr, PHP Admin, XAMPP, DataStax Cassandra.*

**Client Dr. Reddy's Laboratories, India Aug 2016 to Sep 2018**

**Data Engineer**

Responsibilities:

* Responsible for building scalable distributed data solutions using Hadoop and migrate legacy applications to Hadoop.
* Scheduled automated tasks with Oozie for loading data into HDFS through Sqoop and pre-processing the data with Pig and Hive.
* Worked on scalable distributed computing systems, software architecture, data structures and algorithms using Hadoop, Apache Spark and Apache Storm etc.
* Ingested streaming data into Hadoop using Spark, Storm Framework and Scala.
* Delivered real-time experience and analyzed massive amounts of data from multiple sources to calculate real-time ETA using Confluent Kafka event streaming.
* Developed the technical strategy of using Apache Spark on Apache Mesos as a next generation, Big Data and "Fast Data" (Streaming) platform.
* Implemented Flume, Spark framework for real time data processing.
* Developed simple to complex Map Reduce jobs using Hive and Pig for analyzing the data.
* Optimized Map/Reduce Jobs to use HDFS efficiently by using various compression mechanisms.
* Developed big data ingestion framework to process multi TB data including data quality checks, and transformation, and stored as efficient storage formats like parquet and loaded into Amazon S3 using Spark Scala API and Spark.
* Wrote the Spark code in Scala to connect to Hbase and read/write data to the HBase table.
* Extracted data from different databases and to copy into HDFS using Sqoop and has expertise in using compression techniques to optimize the data storage.
* Implemented Kafka producers to create custom partitions, configured brokers, and implemented High-level consumers to implement the data platform.
* Worked on cloud computing infrastructure (e.g. Amazon Web Services EC2) and considerations for scalable, distributed systems
* Created the Spark Streaming code to take the source files as input.
* Used Oozie workflow to automate all the jobs.
* Exported the analyzed data into relational databases using Sqoop for visualization and to generate reports for the BI team.
* Developed spark programs using Scala, involved in creating Spark SQL Queries and Developed Oozie workflow for spark jobs
* Built analytics for structured and unstructured data and managed large data ingestion by using Avro, Flume, Thrift, Kafka, and Sqoop.
* Developed Pig UDFs to know the customer behavior and Pig Latin scripts for processing the data in Hadoop.
* Copied the data from HDFS to MongoDB using pig/Hive/Map reduce scripts and visualized the streaming processed data in Tableau dashboard.
* Continuously monitored and managed the Hadoop Cluster using Cloudera Manager.

**Environments**: *Python, Big data, Hadoop, HBase, Hive, Spark, Pyspark, Cloudera, Kafka, Sqoop, Jenkins, Unix Shell scripting, GitHub, SQL, Tableau, Power BI.*

**Client Advancesoft Tech, Bangalore June 2014 to May 2016**

**Hadoop/ Java Developer,**

**Responsibilities:**

* Installed and configured Hadoop MapReduce, HDFS and developed multiple MapReduce jobs in Java for data cleansing and preprocessing.
* Involved in loading data from UNIX file system to HDFS. Installed and configured Hive and written Hive UDFs.
* Importing and exporting data into HDFS and Hive using Sqoop
* Used Cassandra CQL and Java APIs to retrieve data from Cassandra table.
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshooting, manage and review data backups, manage and review Hadoop log files.
* Worked hands on with ETL process using Informatica.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS.
* Worked extensively on creating tables, views, and SQL queries in MS SQL Server.
* Worked with internal architects and assisting in the development of current and target state data architectures.
* Coordinate with the business users in providing appropriate, effective, and efficient way to design the new
* Partner with technical and non-technical resources across the business to leverage their support and integrate our efforts.
* Partner with infrastructure and platform teams to configure, tune tools, automate tasks and guide the evolution of internal big data ecosystem; serve as a bridge between data scientists and infrastructure/platform teams.
* Write Python scripts to parse JSON documents and load the data in database.
* Generating various capacity planning reports (graphical) using Python packages like Numpy, matplotlib.
* Used Visualization tools such as Power view for excel, Tableau for visualizing and generating reports.
* Worked on the NoSQL databases HBase and mongo DB.
* Performed Exploratory Data Analysis, trying to find trends and clusters.
* Worked on data that was a combination of unstructured and structured data from multiple sources and automated

the cleaning using Python scripts.

* Extensively performed large data read/writes to and from csv and excel files using pandas.

**Environment:** *HDFS, Map Reduce, Pig, Mesos, AWS Hive, Sqoop, Scala, Flume, Mahout, HBase, Spark, SPARK SQL, Yarn, Java, Maven, Git, Cloudera, MongoDB, Eclipse and Shell Scripting.*