**Abhishek Sr. Big Data Engineer**

**Email:**emailabhi75@gmail.com **Visa:** Green card

**Contact:** +1 469 846 8345

**PROFESSIONAL SUMMARY:**

* Around **9+ years** of IT experience in AWS, Cloud, Analysis, design, development, implementation, Big Data Warehouse, NOSQL, GCP, Hadoop, maintenance and support with experience in developing strategic methods for deploying big data technologies to efficiently solve Big Data processing requirement
* Having Hands-on experience in Talend
* Strong experience in BigData using HADOOP framework and related technologies such as HDFS, Teradata, HBASE, MapReduce, PySpark, Spark, Snowflake, Python, HBase, HIVE, PIG, FLUME, OOZIE, SQOOP, and ZOOKEEPER
* Expertise in Extraction Transformation Loading (ETL) process using Informatica consisting of administration, data transformation, data sourcing, mapping, conversion and loading.
* Performance tuning in Informatica Mappings by identifying the bottlenecks and Implemented effective transformation Logic
* Having excellent experience in Data Pipeline and ETL Tools such as Apache Airflow, AWS Data Pipeline, AWS Glue, Talend
* Expert level experience in designing, building and managing applications to process large amounts of data in a Hadoop/DevOps (GCP) ecosystem.
* Extensive experience with performance tuning applications on Hadoop/GCP and configuring Hadoop/GCP systems to maximize performance
* Experience in data analysis using HIVE, PIG LATIN, HBASE and custom Map Reduce programs in Java
* Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, NOSQL, Data Bricks and Azure SQL Data warehouse and Controlling and granting database access and Migrating On premise databases to Azure Data lake store using Azure Data factory
* Experience in writing custom UDFs in JAVA and SCALA for HIVE and PIG to extend the functionality.
* Experience with Cloudera and Horton works distributions.
* Around 3-year 2 months’ experience on SPARK, SCALA, Python, HBASE and KAFKA.
* Developed analytical components using KAFKA, SCALA, SPARK, HBASE and SPARK STREAM.
* Experience in working with Flume to load the log data from multiple sources directly into HDFS
* Excellent experience with data processes, data quality reports
* Performed Hadoop backup Strategy to take the backup of hive, HDFS, HBase, oozie etc.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems (RDBMS) and from RDBMS to HDFS
* Extensive experience in Strategic development of a Data Warehouse and in Performing Data Analysis and Data Mapping from a Operational Data Store to a Enterprise Data Warehouse
* Involved in creating HDINSIGHT cluster in MICROSOFT AZURE PORTAL also created EVENTSHUB and AZURE SQL DATABASES.
* Worked on a clustered Hadoop for Windows Azure using HDInsight and HORTON WORKS Data Platform for Windows.
* Built real time pipeline for streaming data using EVENTSHUB/MICROSOFT AZURE Queue and SPARK STREAMING
* Loaded the aggregated data into HBase for reporting purpose!
* Read the data from HBase to Spark to perform Join on different tables.
* 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 workflow to invoke the Jobs in predefined Interval.
* Expert in scheduling Oozie coordinator based on input data events it starts Oozie workflow when input data is available.
* 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.
* Experienced in managing Hadoop Cluster using HORTONWORKS AMBARI
* 2+ years hands on experience in design, develop and maintain BI Solutions using, Amazon S3 & Redshift.
* Strong Experience in the Business Requirements Analysis, design, conversion of business requirements into high level functional specifications.
* Ability to interact with various groups ranging from Business Users, Business IM teams, Analysts. Had long lasting relationships with Clients in all previous engagements.

**PROFESSIONAL EXPERIENCE**

**Mass Mutual, Springfield, MA April 2021 to Present**

**Big Data Engineer**

**Responsibilities:**

* Wrote Programs in Spark using Scala and Python for Data quality check.
* Worked on Big Data infrastructure for batch processing and real time processing. Built scalable distributed data solutions using Hadoop.
* Written transformations and actions on data frames used Spark SQL on data frames to access hive tables into spark for faster processing of data.
* Imported and exported terabytes of data using Sqoop and real time data using Flume and Kafka.
* Created various hive external tables, staging tables and joined the tables as per the requirement.
* Implemented static Partitioning, Dynamic partitioning and Bucketing in Hive using internal and external table.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Used Hive to do transformations, joins, filter and some pre-aggregations after storing the data to HDFS.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into Cassandra.
* Worked extensively with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).
* Used Scala to convert Hive/SQL queries into RDD transformations in Apache Spark.
* Implemented the workflows using Apache Oozie framework to automate tasks. Used Zookeeper to co-ordinate cluster services.
* Have used Enterprise Data Warehouse (EDW) architecture and various data modeling concepts like star schema, snowflake schema in the project.
* Configured deployed and maintained multi-node Dev and Test Kafka Clusters and implemented data ingestion and handling clusters in real time processing using Kafka.
* Performed various benchmarking steps to optimize the performance of spark jobs and thus improve the overall processing.
* Developed multiple POCs using Pyspark and deployed on the Yarn cluster, compared the performance of Spark, with Hive and SQL/Teradata.
* Developed code in reading multiple data formats on HDFS using Pyspark.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive and involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run MapReduce jobs in the backend.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Designed ETL workflows on Tableau, Deployed data from various sources to HDFS and generated reports using Tableau.
* Worked with SCRUM team in delivering agreed user stories on time for every Sprint.

**Environment**: Hadoop 2.8, MapReduce, HDFS, Yarn, Hive 2.1, Sqoop 1.1, Cassandra 2.7, Oozie, Spark, Scala, Python, AWS, Flume 1.4, Kafka, Tableau, Linux, Shell Scripting.

**AT&T, Dallas, TX Feb 2019 to March 2021**

**Big Data Engineer**

**Responsibilities:**

* Developed Fully automated Configuration driven data pipeline using Spark, Hive, HDFS, SQL Server, oozie and azure File storage to load the client data into Mirror databases.
* Performed necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
* Explored the usage of Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL and Spark Yarn
* Experienced in working with spark eco system using Spark SQL and Scala queries on different formats like Text file, CSV file, transformation in GCP
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Expertise On optimizing spark Jobs when dealing with Huge joins and data Skew.
* Written oozie workflow to invoke the Jobs in predefined Intervals.
* Expert in scheduling Oozie coordinator based on input data events it starts Oozie workflow when input data is available.
* Developed Robust application with the help of Drools-engine In-order to execute complex rules on top of customer data.
* Pretty Good knowledge On the Hortonworks administration and security things such as Apache Ranger, Knox Gateway, High Availability.
* Performed Hadoop backup Strategy to take the backup of hive, HDFS, HBase, oozie etc.
* Used Hive to analyze the Partitioned and Bucketed data and compute various metrics for reporting.
* 2+ years hands on experience in design, develop and maintain BI Solutions using, Amazon S3 & Redshift.
* Strong Experience in the Business Requirements Analysis, design, conversion of business requirements into high level functional specifications.
* Strong knowledge multiple cloud technologies including, EC2, S3, RedShift, VPC, EBS, ELB, EMR, DynamoDB, Lambda, Route 53.
* Ability to interact with various groups ranging from Business Users, Business IM teams, Analysts. Had long lasting relationships with Clients in all previous engagements.
* Expertise in Extraction Transformation Loading (ETL) process using Informatica consisting of administration, data transformation, data sourcing, mapping, conversion and loading.
* Performance tuning in Informatica Mappings by identifying the bottlenecks and Implemented effective transformation Logic.
* Performed hive performance tuning aspects like Map join, cost-based optimization and column level statistics.
* 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.
* Experience in creating UDF's, UDAF's for Hive and Pig.
* Extensively used Pig for data cleansing and HIVE queries for the analysts.
* Performed Benchmark between Hive and Spark SQL.

**Environment**: Azure, AWS, Spark, Hive, PySpark, GCP, Spark SQL, Kafka, Horton Works, JBoss Drools, Hive, Pig, Oozie, HBase, Python, Scala, Teradata, Snowflake, Big Data Warehouse,Maven, Jupiter Notebook, Visual Studio, Unix Shell Scripting.

**Trimble TX Jan 2016 to Dec 2018**

**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.
* Developing UDFs in java for hive and pig and worked on reading multiple data formats on HDFS using Scala.
* 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.
* Developed analytical component using Scala, Spark and Spark Streaming.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis
* 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.
* 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 Hive tables.
* 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 & created Hive views on data in HDFS using Spark.
* Expert in Troubleshooting MapReduce Jobs.
* 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 web Methods 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 and integrating HIVE with existing applications.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Involved in creating ETL flow using Pig, loading with data and writing Pig Latin queries which will run internally in Map Reduce way.
* Involved in writing Unix/Linux Shell Scripting for scheduling jobs and for writing pig scripts and hive QL.
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run 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, Meta store management.
* Developed and designed system to collect data from multiple portal using Kafka and then process it using spark.

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

**Oathmap India Jan 2012 to Dec 2014**

**Java developer**

**Responsibilities:**

* Designed and developed user interfaces using HTML, JSP, and Struts tags.
* Experienced in developing applications using all Java/J2EE technologies like Servlets, JSP, EJB, JDBC, etc.
* Validating the views using validator plug-in in Struts Framework.
* Writing test cases using JUNIT, doing test-first development.
* Writing build files using ANT. Used Maven in conjunction with ANT to manage build files
* Used Hibernate for data persistence and interaction with the database.
* Involved in developing the Struts Action Classes.
* Develop test cases for Unit testing and sanity testing

**Environment:** Oracle, JDK, Struts, Hibernate, Tomcat, Windows 2000