**Karthik Karamchedu**

**Sr. Big Data Engineer**

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

**+1 (469) 846-8345**

**PROFESSIONAL SUMMARY:**

* Over 12 years of IT experience in analysis, design and development using Big Data Ecosystem (Pig, Hive, Impala and Spark, NIFI, Scala), GCP, Java and J2EE and maintenance of Data Warehouse applications using Hadoop Eco-Systems
* Expert in 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, AWS EC2, S3, Auto Scaling, IAM, Lambda, Elastic Load Balancing and other services of AWS.
* Progressive experience in all phases of the iterative Software Development Life Cycle (SDLC)
* Experience in working in environments using Agile (SCRUM), RUP and Test-Driven development methodologies
* Passionate and highly skilled in performing complex data analysis, data mining, predictive model, designing and developing data warehouses
* 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
* 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
* Good understanding of NoSQL Database and hands on work experience in writing application on No SQL database which is MongoDB.
* Strong knowledge in using MapReduce programming model for analyzing the data stored in Hadoop.
* Extensive experience in installing, configuring and using Big Data ecosystem components like MapReduce, HDFS, Sqoop, Pig, Impala & Spark
* Expertise in using J2EE application servers such as IBM Web Sphere, JBoss and web servers like Apache Tomcat
* Created dataflow between SQLServer and Hadoopclusters using ApacheNifi
* Good knowledge on spark components like Spark SQL, MLLib, Spark Streaming and GraphX,
* Proficient in configuring Zookeeper, Cassandra & Flume to the existing Hadoop cluster.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems.
* Expert in configuring and administering the Hadoop Cluster using major Hadoop Distributions like Apache Hadoop and Cloudera
* Experience in using Cloudera Manager for installation and management of single-node and multi-node Hadoop cluster (CDH4&CDH5).
* Experience in Data load management, importing & exporting data using SQOOP & FLUME.
* Experience in analyzing data using Hive, Pig and custom MR programs in Java
* Worked extensively on building Nifi data pipelines in docker container environment in development phase.
* Worked with DevOps team to ClusterizeNIFI Pipeline on EC2 nodes integrated with Spark, Kafka, Postgres running on other instances using SSL handshakes in QA and Production Environments
* Experience in analyzing data using HiveQL, PIG Latin and custom Map Reduce programs in JAVA.
* Experience in different Hadoop distributions like Cloudera (CDH3 & CDH4) and Horton Works Distributions (HDP).
* Good knowledge on using Sqoop to import data into HDFS from RDBMS and vice-versa.
* Experience in working with Version Control Tools like Rational Team Concert, Harvest, Clear Case, SVN, and Git-hub.
* Excellent technical and analytical skills with clear understanding of design goals of ER modeling for OLTP and dimension modeling for OLAP.
* Experience to develop enterprise applications with MVC architecture with application servers and Web.
* Strong experience in analyzing large amounts of data sets writing PySpark scripts and Hive queries.
* Good experience in defining the XML schemas and in working with XML parsers to read and validate the data held in XML documents.
* Hands-on experience in writing Pig Latin scripts, working with grunt shells and job scheduling with Oozie.
* Involve in moving all log files generated from various sources to HDFS and Spark for further processing.
* Experience in Object Oriented Analysis, Design (OOAD) and development of software using UML Methodology
* Experience on R and Python for statistical computing. Also experience with MLlib (Spark), Matlab, Excel, Minitab, SPSS, and SAS
* Experience in configuring various topologies in storm to ingest and process data on the fly from multiple sources and aggregate into central repository Hadoop.
* Experience developing iterative algorithms using Spark Streaming in Scala and Python to builds near real-time dashboards.
* Extensive experience working with Spark tools like RDD transformations, Spark MLlib and SPARQL.
* Experienced in moving data from different sources using Kafka producers, consumers and pre-process data using Storm topologies.
* Experienced in migrating ETL transformations using Pig Latin Scripts, transformations, join operations.
* Good understanding of MPP databases such as HPVertica, Greenplum and Impala.
* Hands on experience in implementing Sequence files, Combiners, Counters, Dynamic Partitions and Bucketing for best practice and performance improvement.
* Highly Knowledgeable in streaming data from different data sources like Log files, JMS, applications sources into HDFS using Flume sources.
* Knowledge of data warehousing and ETL tools like Informatica, and Talend.
* Experience with Testing MapReduce programs using MR Unit, Junit.
* Experience working on Version control tools like SVN and Git revision control systems such as GitHub and JIRA/MINGLE to track issues and crucible for code reviews.
* Worked on various Tools and IDEs like Eclipse, IBM Rational, Visio, Apache Ant-Build Tool, MS-Office, PLSQL Developer, SQL\*Plus.
* Experience in different application servers like JBoss/Tomcat, WebLogic, IBM WebSphere.
* Experience in working with Onsite-Offshore model.
* Implemented logging framework - ELK stack (Elastic Search, Logstash & Kibana) on AWS.

**TECHNICAL SKILLS:**

* **Big Data Frameworks**: Hadoop, Spark, NIFI, Scala, Hive, Kafka, AWS, Cassandra, HBase, Flume, Pig, Sqoop, Map Reduce, Cloudera, Mongo DB, Azure SQL Data Warehouse
* **Big data distribution**: Cloudera, Amazon EMR
* **Programming languages**: Core Java, Scala, Python, SQL, Shell Scripting
* **Operating Systems**: Windows, Linux (Ubuntu)
* **Databases**: Oracle, SQL Server
* **Designing Tools**: Eclipse
* **Java Technologies**: JSP, Servlets, Junit, Spring, Hibernate
* **Web Technologies**: XML, HTML, JavaScript, JVM, JQuery, JSON
* **Linux Experience**: System Administration Tools, Puppet, Apache
* **Web Services**: Web Service (RESTful and SOAP)
* **Frame Works**: Jakarta Struts 1.x, Spring 2.x
* **Development methodologies**: Agile, Waterfall
* **Logging Tools**: Log4j
* **Application / Web Servers**: Cherrypy, Apache Tomcat, WebSphere
* **Messaging Services**: ActiveMQ, Kafka, JMS
* **Version Tools**: Git, SVN and CVS
* **Analytics**: Tableau, SPSS, SAS EM and SAS.

**PROFESSIONAL EXPERIENCE:**

**Sr. Big Data Engineer AUG 2020 to SEP 2022**

**Mass Mutual, Springfield, MA**

**Responsibilities:**

* As a Sr. Big Data Developer worked on Hadoop eco-systems including Hive, MongoDB, Zookeeper, Spark Streaming with MapR distribution
* Designed and developed data mapping and transformation scripts to support and promote data warehouse development, structural changes of multiple RDBMS and data analytics efforts as well as design effective ETL logic and code as required
* Developed Big Data solutions focused on pattern matching and predictive modeling.
* Implemented Security in Web Applications using Azure and deployed Web Applications to Azure.
* Worked on analyzing Hadoop cluster and different big data analytic tools including Pig, HBase database and Sqoop
* Implemented various Azure platforms such as Azure SQL Database, Azure SQL Data Warehouse, Azure Analysis Services, HD Insight, Azure Data Lake and Data Factory
* Participated in all aspects of Software Development Life Cycle (SDLC) and Production troubleshooting, Software testing using Standard Test Tool
* Experienced in working with spark eco system using Spark SQLand Scala queries on different formats like Text file, CSV file,transformation in GCP
* Involved in Agile methodologies, daily scrum meetings, spring planning.
* Involved in writing Spark applications using Scala to perform various data cleansing, validation, transformation and summarization activities according to the requirement.
* Implemented solutions utilizing Advanced AWS Components: EMR, EC2, etc integrated with Big Data/HadoopDistributionFrameworks: Zookeeper, Yarn, Spark, Scala, NiFi etc.
* Developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Involved in identifying job dependencies to design workflow for Oozie and YARN resource management.
* Designed solution for various system components using Microsoft Azure.
* Worked on data using Sqoop from HDFS to Relational Database Systems and vice-versa. Maintaining and troubleshooting
* Explored with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, pair RDD.
* Created Hive Tables, loaded claims data from Oracle using Sqoop and loaded the processed data into target database.
* Exported data from HDFS to RDBMS via Sqoop for Business Intelligence, visualization and user report generation.
* Developed ApacheNifi flows dealing with various kinds of data formats such as XML, JSON, and Avro.
* Worked on importing data from HDFS to MYSQL database and vice-versa using SQOOP.
* Configured Hive Meta store with MySQL, which stores the metadata for Hive tables.
* Performed data analytics in Hive and then exported those metrics back to Oracle Database using Sqoop.
* Upgraded the Hadoop Cluster from CDH3 to CDH4, setting up High Availability Cluster and integrating Hive with existing applications.
* Worked on NoSQL support enterprise production and loading data into HBase using Impala and Sqoop.
* Performed multiple MapReduce jobs in Pig and Hive for data cleaning and pre-processing.
* Handled importing of data from various data sources, performed transformations using Hive, PIG, and loaded data into HDFS.
* Proactively involved in ongoing maintenance, support and improvements in Hadoop cluster.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Used Hive to analyze data ingested into HBase by using Hive-HBase integration and compute various metrics for reporting on the dashboard.
* Worked on analyzing, writing Hadoop MapReduce jobs using JavaAPI, Pig and hive.
* Analyzed large amounts of data sets using HBase to aggregate and report on it.
* Developed reports, dashboards using Tableau for quick reviews to be presented to business.
* Worked on configuring and managing disaster recovery and backup on Cassandra Data.
* Developed many distributed, transactional, portable applications using Enterprise JavaBeans (EJB) architecture for Java 2 Enterprise Edition (J2EE) platform.
* Used Cloudera Manager for installation and management of Hadoop Cluster.
* Developed data pipeline using Flume, Sqoop, Pig and Java MapReduce to ingest customer behavioral data and financial histories into HDFS for analysis.
* Worked on MongoDB, HBase databases which differ from classic relational databases
* Involved in converting HiveQL into Spark transformations using Spark RDD and through Scala programming.
* Used Hive to perform data validation on the data ingested using Sqoop and cleansed the data.
* Developed several business services using Java RESTful Web Services using Spring MVC framework.

**Environment**: Hadoop 3.0, Oozie 4.3, GCP, Zookeeper 3.4, Cassandra 3.0, Sqoop 1.4, Apache NiFi 1.4, ETL, Azure, Hive 2.3, HBase 1.4, Pig 0.17, HDFS 3.1, Flume 1.8, Tableau, GIT, Kafka 1.1, MapReduce, JSON, AVRO, Teradata, Maven, SOAP.

**Bigdata AWS Cloud Engineer, August 2018 to JULY 2020**

**AT&T, Dallas, TX**

**Responsibilities:**

* Responsible for building scalable distributed data solutions using Hadoop.
* Involved in gathering the business requirements from the Business Partners and Subject Matter Experts.
* Migrated MapReduce jobs to Spark jobs to achieve better performance.
* Interacted with the stake-holders and gather requirements and business artifacts based on Agile SCRUM methodology
* Migrated on premise database structure to Redshift data warehouse.
* Extracted Real time feed using Kafka and Spark streaming and convert it to RDD and process data in the form of Data Frame.
* Worked on Kafka and REST API to collect and load the data on Hadoop file system also used Sqoop to load the data from relational databases
* Migrated their Big Data Platform from on-premise Hadoop to Google Cloud Platform (GCP)to one of the projects that we are working
* Wrote Spark-Streaming applications to consume the data from Kafka topics and write the processed streams to HBase
* Designed and built end-to-end Data Warehouse infrastructure from the ground up on Redshift for large scale data handling
* Open SSH tunnel to GCP-Google DataProc to access to yarn manager to monitor spark jobs
* Manage and support of enterprise Data Warehouse operation, big data advanced predictive application development using Cloudera & Horton works HDP
* Used Spark Data Frames Operations to perform required Validations in the data and to perform analytics on the Hive data.
* Developed Apache Spark applications by using spark for data processing from various streaming sources.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs.
* Imported data from AWS S3 and into spark RDD and performed transformations and actions on RDD.
* Implemented Apache Nifi flow topologies to perform cleansing operations before moving data into HDFS.
* Involved in migrating MapReduce jobs into RDD (Resilient data distributions) and create Spark jobs for better performance.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Developed the batch scripts to fetch the data from AWS S3 storage and do required transformations in Scala using Spark framework.
* Developed Scala scripts, UDF using both Data frames/SQL and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop.
* Involved in executing various Oozie workflows and automating parallel Hadoop MapReduce jobs.
* Involved in transforming data from legacy tables to HDFS and Hive tables using Sqoop.
* Implemented Spark using and Spark SQL for faster testing and processing of data responsible to manage data from different sources Scala.
* 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).
* Developed Oozie Bundles to Schedule Pig, Sqoop and Hive jobs to create data pipelines.
* Experienced in using ORC, Avro, Parquet, RCFile and JSON file formats and developed UDFs using Hive and Pig.
* Used PIG to perform data validation on the data ingested using Sqoop and Flume and the cleansed data set is pushed into MongoDB.
* Implemented multiple MapReduce Jobs in java for data cleansing and pre-processing.
* Developed Hive queries to do analysis of the data and to generate the end reports to be used by business users.
* Import the data from different sources like HDFS/HBase into Spark RDD and developed a data pipeline using Kafka and Storm to store data into HDFS.
* Used Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scala and NoSQL databases such as HBase and Cassandra.
* Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS, HBase and Elastic Search.
* Explored MLlib algorithms in Spark to understand the possible Machine Learning functionalities that can be used for our use case
* Worked with teams in setting up AWS EC2 instances by using different AWS services like S3, EBS, Elastic Load Balancer, and Auto scaling groups, VPC subnets and CloudWatch.
* Imported and exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team.
* Created and maintained various Shell and Python scripts for automating various processes and optimized MapReduce code, pig scripts and performance tuning and analysis.

**Environment**: Hadoop 3.0, Spark, Hive 2.3, Agile, MapReduce, Kafka 1.1, HBase 1.4, HDFS 3.1, Sqoop 1.4, Scala, AWS, RDBMS, Oozie, Pig 0.17, Sqoop, Cassandra 3.11, NoSQL, NIFI, GCP, Elastic Search, Java,

**Big Data Engineer March 2015 to NOV 2017**

**Cigna Health Insurance ‎Wilmington, Delaware**

**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.

**GE Power, Chicago, IL Nov 2011 – Feb 2015**

**Big Data Engineer**

**Responsibilities:-**

* Developed a data pipeline to ingest customer behavioral data and financial histories into Hadoop cluster for analysis.
* 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.
* Involved in all phases of Installation and upgradation of Hadoop big data platform. Implementing security for Hadoop big data platform
* Designed the sequence diagrams to depict the data flow into Hadoop.
* Involved in importing and exporting data between HDFS and Relational Systems like Oracle, MySQL and DB2 using Sqoop.
* Setup best practices for monitoring. Analyze Hardware, Software requirements for the projects
* Helped Application and Operations team to troubleshoot the performance issues
* Implemented Partitioning, Dynamic Partitions and bucketing in HIVE for efficient data access.
* Created final tables in Parquet format. Use of Impala to query and manage Parquet tables.
* Implemented data Ingestion and handling clusters in real time processing using Apache Kafka.
* Involve in creating Hive tables, loading with data and writing Hive queries
* Collected data using Spark from AWS S3 bucket in near-real-time and performs 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.
* Developed Spark Code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Involved in converting Hive/SQL queries into Spark Transformations using Spark RDDs and Scala.
* Worked on the Spark SQL and Spark Streaming modules of Spark and used Scala and Python to write code for all Spark use cases.
* Explored the Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark-Context, Spark-SQL, Data Frame and Pair RDD's.
* Migrated historical data to S3 and developed a reliable mechanism for processing the incremental updates.
* Used Oozie workflow engine to manage independent Hadoop jobs and to automate several types of Hadoop such as java MapReduce, Hive and Sqoop as well as system specific jobs
* Used to monitor and debug Hadoop jobs/applications running in production.
* Worked on providing user support and application support on Hadoop infrastructure.
* Worked on evaluating, comparing different tools for test data management with Hadoop
* Supported the testing team on Hadoop Application Testing.

**Environment**: Hadoop, HDFS, Pig, Hive, Spark, MapReduce, Azure,Java, Cloudera CDH 4.6, Hadoop, HDFS, Map Reduce, Hive, Sqoop, Oozie

**Brook Source, Chicago, IL Aug 2010 to Sept 2011**

**Java Developer**

**Responsibilities:**

* 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.
* Created web service components using SOAP, XML and WSDL to receive XML messages and for the application of business logic.
* Involved in configuring web sphere variables, queues, DSs, servers and deploying EAR into Servers.
* Involved in developing the business Logic using Plain Old Java Objects (POJOs) and Session EJBs.
* Developed authentication through LDAP by JNDI.
* Developed and debugged the application using Eclipse IDE.
* Involved in Hibernate mappings, configuration properties set up, creating sessions, transactions and second level cache set up.
* Involved in backing up database & in creating dump files. And also creating DB schemas from dump files. Wrote developer test cases & executed. Prepared corresponding scope & traceability matrix.
* Implemented JUnit and JAD for debugging and to develop test cases for all the modules.
* Hands-on experience of Sun One Application Server, Web logic Application Server, Web Sphere Application Server, Web Sphere Portal Server, and J2EE application deployment technology.

**Environment:** Java multithreading, JDBC, Hibernate, Struts, Collections, Maven, Subversion, JUnit, SQL language, Struts, JSP, SOAP, Servlets, Spring, Hibernate, Junit, Oracle, XML, Putty and Eclipse.

**EDUCATION: Master’s in Electrical engineering 2005**