**Kalyan G**

(816) 286-4992 •  [giraboina.4992@gmail.com](mailto:%20giraboina.4992@gmail.com)

# Professional summary:

* Over 7 years of professional IT experience which includes over 3 years of experience in **Big data ecosystem related technologies** and continuous work experience in **Java**.
* Experience in architecting, designing, installation, configuration and management of **Apache Hadoop Clusters, MapR, Hortonworks & Cloudera Hadoop Distribution**.
* Experience in managing the Hadoop infrastructure with **Cloudera Manager**.
* Experienced in Big Data Ecosystem with **Hadoop, HDFS, MapReduce, Pig, Hive, HBase, Impala, Sqoop, Flume, Kafka, Oozie, Spark, PySpark and Spark Streaming**.
* Proficient in **Java, Python, and Scala** in **Apache Spark**.
* Strong experience with **Pig, Hive, Impala, MapReduce** in Hadoop Ecosystem.
* Experience in setting up and maintaining **Hadoop cluster** running **HDFS** and **MapReduce** on **YARN**.
* Strong Database Experience on **RDBMS (Oracle, MySQL) with PL/SQL** programming skills in creating **Packages, Stored Procedures, Functions, Triggers & Cursors**.
* Extensive familiarity with **SQL, Oracle** and **MySQL** database management.
* Good exposure on usage of **NoSQL** databases column-oriented **HBase**, **Cassandra** and **MongoDB** **(Document Based DB).**
* Experience in importing and exporting data using **Sqoop** from **HDFS** to **Relational Database Systems (RDBMS)** and vice-versa.
* Exposure on the **HBase** distributed database and the **ZooKeeper** distributed configuration service.
* Handling and further processing **schema oriented** and **non-schema** oriented data.
* Using **Spark Streaming**, ingested data from **Kafka**, **TCP sockets** and processed with high-level functions such as **map, reduce, join and window**, and processed data is pushed out into file systems and/or databases.
* Involved in loading and transforming large sets of **structured data** from router location to EDW using a **Nifi** data pipeline flow.
* Designed Data flow to pull the data from **Rest API** using **Apache Nifi** with **SSL** context configuration enabled.
* Experience in implementing unified data platforms using **Kafka producers/ consumers**.
* Developed **Kafka** **consumer** to receive and store real time data from **Kafka** to **Amazon S3.**
* Experience in building **Scala**, **Spark SQL** and **MLlib** libraries along with **Kafka** and other tools as per requirement then deployed on the **Yarn cluster**.
* Using unified analytical tool; **Databricks (PySpark, SparkSQL, Scala, DataFrames, Datasets etc.)**
* Experienced with distributions including **Amazon EMR 4.x** and **Hortonworks HDP 2.2**
* Deployed Web Application on **Amazon EC2.**
* Used **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).**
* Working experience on **C****ore java** technology, which includes **Class-design**, **Multithreading**, **I/O&JDBC**, **Collections**, **Localization**, ability to develop new **API** for different projects.
* Excellent proficiency in **Tomcat Apache** and IIS web servers
* Strong hands-on development experience with **Java, J2EE (Servlets, JSP, Java Beans, EJB, JDBC, JMS, Web Services)** and related technologies.
* Strong experience developing **J2EE** applications, Enterprise Applications using **Java**, **J2EE**, **Spring Framework, Hibernate, Web Services (SOAP and RESTFUL) and Junit Testing**.
* Experience in building, deploying and integrating applications in Application Servers with **ANT, Maven** and **Gradle**.
* Experience working in **MVC** framework using Spring Framework including **Spring MVC, Spring IOC, Spring JDBC**.
* Experience with front end technologies **HTML (5), CSS, JavaScript, XML and jQuery.** Worked extensively on different databases **Oracle, MySQL** and have good database programming experience with **SQL**.
* Experienced with the entire **Software Development Lifecycle (SDLC)** of applications: gathering requirements, analysis, conceptual and detail design, development, verification and testing.
* Experience in using IDE tools such as **Visual Studio, NetBeans, and Eclipse** and application servers **WebSphere, WebLogic** and **Tomcat**
* Expertise in all phases of System Development Life Cycle Process (SDLC), **Agile** Software Development, **Scrum** Methodology and Test-Driven Development.
* Used **Tomcat server** for the application development and Utilized **JIRA** for task scheduling.
* Experience in using Version Control tools like **Git, SVN.**
* Experience of application development in different environments like **Windows** and **Linux**.

# Technical Skills:

|  |  |
| --- | --- |
| **Hadoop/Big Data:** | MapReduce, HDFS, Hive 2.3, HBASE 1.2, Sqoop 1.4, Flume 1.8, Scala 2.12, Hadoop 3.0, Spark, Impala, Pig, SparkSQL, Cassandra, Kafka, Oozie, PySpark, YARN, ZooKeeper. |
| **Database Skills:** | SQL-Server, MySQL, SQLite, MongoDB, Oracle |
| **Cloud Technology:** | Amazon Web Services (AWS)-EMR, EC2, S3, CloudFormation, Elastic Search, Microsoft Azure. |
| **Web Tools:** | HTML, XML, Java Script, ODBC, JDBC, Hibernate, JSP, Servlets, Java, Struts, spring, and Avro. |
| **Languages:** | Java, SQL, Shell Scripting, Python, JavaScript, C & C++, jQuery, AJAX, CSS, XML, DOM, SOAP, REST. |
| **IDE and Build Tools:** | Eclipse, Maven, JIRA, Jenkins, ANT, NetBeans, IntelliJ |
| **Version controls:** | SVN, Confluence Version Control Git. |
| **Operating System:** | Windows, Unix, Linux. |
|  |  |

# Professional Experience:

**Client: CenturyLink – Salt Lake City, Utah, USA Sep 2019 – Present**

**Role:** **Spark Developer**

**Responsibilities:**

* Worked in a fast-paced **agile** **development** environment to quickly analyze, develop, and test potential use cases for the business.
* Participated in development/implementation of **Cloudera Hadoop environment.**
* Programming using **Python, core JAVA along with Hadoop framework** utilizing **Cloudera Hadoop Ecosystem** projects (**HDFS, Spark, Sqoop, Hive, HBase, Oozie, Impala, Zookeeper etc.).**
* Experience in managing **Hadoop clusters** using **Cloudera Manager tool**.
* 
* Skilled with Python **parsing, manipulating, and converting data** to and from a wide range of formats (**CSV, json, XML, html**, etc.,)
* Used **Amazon EMR** to create and configure a cluster of **Amazon EC2** instances running **Hadoop**.
* Responsible for developing **data pipeline** with **Amazon AWS** to extract the data from weblogs and store in **HDFS**.
* Maintaining existing **ETL** workflows, data management and data query components.
* Good understanding of NoSQL databases and hands on work experience in writing applications on NoSQL databases like **HBase, Cassandra and MongoDB**.
* Implemented **Nifi** flow topologies to perform **cleansing** operations before moving data into **HDFS**.
* Worked with the **Apache Nifi** flow to perform the conversion of Raw data into **ORC**.
* Involved in loading and transforming large sets of **structured data** from router location to EDW using a **Nifi** data pipeline flow.
* Designed Data flow to pull the data from **Rest API** using **Apache Nifi** with **SSL** context configuration enabled.
* Implemented the **Cassandra** and manage of the other tools to process observed running on over **YARN**.
* Created a POC for the demonstration of retrieving the **JSON** data by calling **Rest** service and converting into **CSV** by creating data flow and loading into **HDFS**.
* Designed and created **Hive** external tables using shared **meta-store** instead of derby with partitioning, dynamic partitioning and buckets.
* Implemented schema extraction for **Parquet** and **Avro** file Formats in **Hive**.
* Used unified analytical tool; **Databricks (PySpark, SparkSQL, Scala, DataFrames, Datasets etc.)**
* Implemented **Spark** using **Scala** and **SparkSQL** for faster testing and processing of data.
* Developing **Spark** programs using **Scala API's** to compare the performance of **Spark** with **Hive** and **SQL**.
* **Extract, transform, and load (ETL)** data from multiple federated data sources (**JSON**, **relational database**, etc.) with **DataFrames** in **Spark**.
* Utilized **SparkSQL** to extract and process data by parsing using **Datasets** or **RDDs** in **HiveContext**, with transformations and actions (map, flatMap, filter, reduce, reduceByKey).
* Analyzed the **SQL** scripts and designed the solution to implement using **PySpark.**
* Extended the capabilities of **DataFrames** using **User Defined Functions** in **Python** and **Scala**.
* Interaction with **Spark Shell** using **Python API- PySpark.**
* Used **Spark API** over **Cloudera Hadoop YARN** to perform analytics on data in **Hive**.
* **Spark Streaming** is used to process **Kafka's** real-time data, and the processed data is sent out to file systems and/or databases.
* Using **Spark Streaming**, ingested data from **TCP sockets** and processed with high-level functions such as **map, reduce, join and window**, and processed data is pushed out into file systems and/or databases.
* Divided the data stream into batches called **DStreams** and processed using Spark APIs and returned the results in batches for batch streaming of data.
* 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**.
* Using **Kafka** to build real-time **data pipelines and streaming** applications, **publish and subscribe** to message queue (**Topic**), o Store streams of records in a fault-tolerant durable way, and process streams of records as they occur.

**Environment**: Hadoop 2.7.7, HDFS 2.7.7, Spark 2.1, MapReduce 2.9.1, Hive 2.3, Kafka 0.8.2.X, HBase, Scala 2.12.8, AWS, Python 3.7, Java 8, JSON, SQL Scripting and Linux Shell Scripting, Avro, Parquet, **Cloudera (CHD 5.X).**

**Client: Deutsche Bank - McLean, VA, USA Sep 2018 – Sep 2019**

**Role:** **Big Data Engineer**

**Responsibilities:**

* Involved in story-driven **agile** development methodology and actively participated in daily **scrum** meetings.
* Well versed in the installation and management of the **Hadoop** **distribution** of **Hortonworks.**
* Created hive tables on top of **Avro** data in **AWS S3** landing zone. Performing joins on hive tables and loading to standardized zone in **parquet** format.
* Developed generic **shell scripts** to Create **AWS EMR** clusters, submit **Scala Spark** and **hive jobs** on **clusters** and terminating the cluster after completion of the job.
* Integrating **AWS RDS** and **Datadog** agent on **EMR** for configuring external metastore, monitoring and **log streaming**.
* Solid Enterprise **Java** and working knowledge of **Scala** programming languages.
* Load and transform large sets of **structured**, **semi structured** and **unstructured data.**
* Contributed in designing, developing and documenting high quality software for large scale **Hadoop** distributed systems by loading and processing datasets of various file formats like **Avro**, **Parquet** and **JSON**.
* Provided batch processing solution to certain **unstructured** and large volume of data by using **Hadoop Map Reduce** framework.
* Built distributed, scalable, and reliable **data pipelines** that **ingest** and process data at scale using **Hive** and **MapReduce**.
* Created **external hive tables** implemented **dynamic partitioning and bucketing** in hive as part of performance tuning.
* Worked in importing and exporting utility data into **HDFS** and **Hive Metastore** from **RDBMS** (Oracle databases) using **Sqoop**.
* Executed **Hive** queries on **Parquet** tables stored in **Hive** to perform data analysis to meet the business requirements.
* Using impala on top of the Hadoop ecosystem for parallel database processing, low latency queries and **partial data analysis.**
* Worked on POC’s with **Apache Spark** using **Scala** to implement **spark** in project.
* Used the **Spark API** over **Hortonworks Hadoop YARN** to perform data analysis in **Hive**.
* Optimization of existing algorithms in Hadoop using **Spark Hive-Context, Spark-SQL**, **Data Frames** and Pair **RDD's**.
* Explored the integration of **Hive queries** with **Spark-SQL** into **Spark system**.
* Developed **Scala** scripts, UDFFs using both **Data frames/SQL/Data sets** and **RDD/MapReduce in Spark** for Data Aggregation, queries and writing data back into **OLTP** system through **Sqoop**.
* Consumed the data from **Kafka** using **Apache spark**.
* Configured deployed and maintained multi-node Dev and Test **Kafka** Clusters.
* Familiarized with Hadoop Ecosystems such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce programming.**

**Environment**: Hadoop 2.7, HDFS, Spark 2.0, MapReduce 2.9.0, Hive 2.2, Sqoop 1.4.6, Scala 2.11.8, AWS, Java 8, JSON, SQL Scripting and Linux Shell Scripting, **Hortonworks 2.5.6.0**

**Client: Sun Trust Bank - Atlanta, GA May 2016 – Sep 2018**

**Role:** **Hadoop Developer**

**Responsibilities:**

* Installing, migrating and upgrading multiple **MapR** **clusters**.
* Worked on Installation and configuring of **ZooKeeper** to co-ordinate and monitor the cluster resources.
* Involved in loading data from **LINUX** file system to **HDFS.**
* Developing **Java MapReduce** programs for grouping the data and to calculate the algebraic calculations for the **reducer**.
* Knowledge on writing **Map Reduce** code to make **unstructured data** as **structured data** and for inserting data into **HBase** from **HDFS**.
* Migrating **ETL** jobs to **Pig** scripts do **Transformations**, even joins and some pre-aggregations before storing the data into **HDFS**.
* Worked on analyzing **Hadoop** cluster and different big data analytic tools including **Pig** & **Hbase** **Nosql** database.
* Configuring and performance tuning the **sqoop** jobs for importing the input (raw) data from the **data warehouse**.
* Involved in creating **Hive** tables, loading with data and written **Hive UDFs** to extract data from **staging table**.
* Optimized **hive** joins for large tables and developed **map reduce** code for the full outer join of two large tables.
* Developing **hive** and **impala** queries using **partitioning**, **bucketing** and **windowing** functions.
* Created integration between **Hive** and **HBase**.
* Used **Oozie** scheduler to submit workflows.
* Implemented the recurring workflows using **Oozie** to automate the scheduling flow.
* **Importing** and **exporting** data in **HDFS** and **Hive** using **Sqoop**.
* Imported data using **Sqoop** to load data from **MySQL** to **HDFS** on regular basis.
* **Sqoop** jobs, **PIG** and **Hive** scripts were created for data ingestion from **relational databases** to compare with historical data.
* Collecting, aggregating, and moving data from servers to **HDFS** using **Apache Flume**.
* Developed **data pipeline** using **Flume**, **Sqoop**, **Pig** and Java map reduce to ingest customer behavioral data and financial histories into **HDFS** for analysis.
* Used Apache **Flume** to aggregate and move data from **web servers** to **HDFS**.
* Familiarized in Hadoop Ecosystem including **Pig, Hive, HDFS, MapReduce, Sqoop, Storm, Spark, Kafka, Yarn, Oozie, and Zookeeper.**

**Environment**: Java 7, Eclipse Mars, **MapR 4.1.0**, Linux, Hadoop 2.6.2, MapReduce 2.6.2, Hive 1.1.1, Pig 0.15, Centos 6.4, HDFS 2.6.2, MySQL 5.7, Sqoop 1.4.4, Oozie, MongoDB 3.0.5, HBASE, Impala.

**Client: Kony IT Services Pvt. Ltd - Hyderabad, TS, India. May 2014 – May 2016**

**Role:** **Java Developer**

**Responsibilities:**

* Responsible and mentored the team in complete software development lifecycle (SDLC) tasks - design, coding, testing, and documentation using Rational Unified Process (RUP) for analysis and design of application.
* Designed and developed the web-tier using **HTML**, **JSP**’s, **Servlets**, **Struts** and Tiles framework.
* Involved in the development of business module applications using **J2EE** technologies and **JDBC**.
* Used the lightweight container of the **Spring Framework** to provide architectural flexibility for **Inversion of Controller (IOC).**
* **Hibernate** framework is used in persistence layer for mapping an object-oriented domain model to a relational database (**Oracle**).
* Designed the Architecture of the project as per Spring **MVC** Framework. Worked with **Spring Core, Spring AOP, Spring Integration Framework with Hibernate**.
* Used **SQL** statements and procedures to fetch the data from the DB.
* Skilled in writing the **Unix Shell Scripting** and **Python** scripting for automate process.
* Used **Log4J** for logging messages and **Rational Clear Case** for **version** **Control**.

**Environment**: Java 7, J2EE, Spring AOP 4.0, Struts 2.3.14, HTML, CSS, JavaScript, Hibernate 4.2, WebLogic 12.1.2, SQL 2005, ANT 1.9.1, Log4J 2.0, JUnit, XML, JSP, Servlets 3.1, AJAX, Unix, Python 2.6.9, WebSphere Application Server 8.5.

**Client: App labs - Hyderabad, Andhra Pradesh, India. May 2012 - May 2014**

**Role: Jr. Java Developer**

**Responsibilities**:

* Implemented the application using **Spring** **MVC** Framework by implementing **Controller**, Service classes.
* Implemented front-end using **JSP, JSON, HTML, CSS, JavaScript** and **Custom** tags.
* Performed client-side validations using **Java script** and server-side validation using **spring validator framework**.
* Implemented **spring** framework bean factory & application context for bean initialization and configurations.
* Used **Maven** tool for building and deploying the Web applications.
* Configured and created application log files using **Log4j** required to trace application messages
* Wrote project built script by using **Maven** and used **SVN** for version control.

**Environment**: Java/J2EE, JSP, JavaScript, HTML, CSS, JDBC, XML, Servlets, Spring 3.0, Hibernate 4.0, MVC, MySQL 5.5, Log4J, Maven 3.0.2, SVN.