

**Koteswara Singothu**

**Mobile:** +1-469-933-9876 **Email/skype:** [raohadoop2024@gmail.com](mailto:raohadoop2024@gmail.com)

LinkedIn Profile: <https://www.linkedin.com/in/koteswara-rao-singothu-19b32571/> **Professional** **Summary**Having **15+ years** of experience in IT which includes Analysis, Design, Development, implementation, and support of business applications in multiple domains(**Banking**, **Retail** and **Insurance**) using different technologies like **Big Data**, **Cloud** **AWS**, **DevOps**, **ETL tools**, **SQL**, **NoSQL**, **Python**, **Scala**, and **UNIX**.

* Having **8+** years of work experience on **Hadoop Echo system Tools** with hands on experience in installing, configuring, and using ecosystem components like **HDFS, Map Reduce, YARN, Pig, Hive, HBase, Oozie, Impala, TEZ, Sqoop, Kafka, HUE, Nifi and Databricks, Spark** on both **Hortonworks and Cloudera clusters.**
* Having **3+** years of experience in **AWS Cloud** solution development using **S3, Lambda, EC2, EMR, IAM, Athena, Redshift, Glue, Airflow, AWS Data Pipeline,** and **CloudFormation.**
* Having **1+** years of experience in **Snowflake Multi - cluster Warehouse**
* Having **1+** years of experience in **CDP7.X** after collaboration of **Cloudera** and **Hortonworks Hadoop platform.**
* Hands on experience on created data pipeline by using **PySpark** and **AWS** tools like **IAM, LAMBDA, Glue, SNS, CloudWatch** and **Redshift**
* Hands on experience on Unified Data Analytics with **Databricks**, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation and aggregation from multifile file formats for Analyzing and transforming the data to uncover insights into the cluster usage patterns.
* Extensive experience in working on Analyze, Build, Test and Implementation of ETL functionality using **IBM Data Stage, Informatica** and **SSIS** and visualize the data by using BI Report tools like **Tableau, OBIEE, ThoughtSpot** and **MicroStrategy** by using Databases like **Oracle, Teradata, SQL Server, MariaDB, Greenplum** and **MongoDB**.
* Experience in **Spark Core, Spark SQL,** and **Spark Streaming** by using **Python** and **Scala** program to create the batch, Realtime data pipeline to move the heterogenous data into Hadoop Lake or AWS cloud.
* Strong experience in extracting and loading data using complex business logic’s using Spark, Sqoop, PIG from different data sources and built the ETL pipelines to process tera bytes of data daily.
* Experience in migrating the workflow jobs written in **Scalding**, a **MapReduce-based** framework to Spark jobs
* Worked and designed **performance tuning of Spark** applications for adjusting Batch Interval time, level of parallelism for spark jobs and **memory tuning** to get an optimized throughput for a longer running applications to mitigate expenses caused by inefficient use of resources.
* Monitoring Spark containerized applications which are running on Jenkins using IBM monitoring **Grafana**.
* Worked with Spark on an EMR cluster along with other Hadoop applications, and it can also leverage the EMR file system (EMRFS) to directly access data in Amazon S3, Amazon Glue.
* Having experience in Version Control tools like GIT, Bitbucket and DevOps tools like Jenkins and Bamboo and scheduling tools like Crontab, Autosys, Tidal and Control-M
* Experience in writing and running **shell scripts** with SQOOP, HQL, Spark-shell and spark-submit tasks along with file validations and archive data.
* Knowledge on various file formats such as Avro, Parquet, ORC, etc. and on various compression codecs such as GZip, Snappy. ­
* Strong understanding of Software Development Lifecycle (SDLC), and **Agile** methodology).
* Worked in projects based on **Agile-Scrum** Methods.
* Have good experience, excellent communication and interpersonal skills which contribute to timely completion of project deliverable well ahead of schedule.

**Educational:**

* B-Tech (Computer Science Engineering), Sep 2002- July 2006 – Andhra University, India

**Professional Experience:**

* Currently working as a Senior Hadoop Data Engineer since Aug 2019.
* Worked as a Cloud Hadoop Data Engineer in Mphasis Limited from April 2019 to Aug 2019
* Worked as a Cloud Hadoop Data Engineer in Net Logic Solutions from April 2018 to April 2019
* Worked as a Technology Lead in Infosys Pvt Ltd from July 2013 to April 2018.
* Worked as an Associate in Cognizant Technology Solution from July 2009 to Aug 2013.
* Worked as a Programmer Analyst in IP Creations Pvt Ltd from May 2008 to July 2009.

**Certifications:**

CCA Spark and Hadoop Developer: <http://certification.cloudera.com/verify> - **100-024-306**

AWS Certified Solutions Architect–Associate: http://aws.amazon.com/verification --4S0EM7RL11FQQDWZ

AWS Certified Machine Learning – Specialty: http://aws.amazon.com/verification --WE9KBJ4BBJB4QQGF

Coursera Machine Learning: https://coursera.org/share/5c26059eb23ff105d12d2ffb181e987a

Data Science with Math skills: <https://coursera.org/share/8bd20c7f422ee69943a114e49f3c8040>

**Technical Skills:**

|  |  |
| --- | --- |
| Hadoop Platform | Cloudera(CDH), Hortonworks(HDP), CDP |
| Hadoop Tools | HDFS, Hive, PIG, SQOOP, HBase, Oozie, Tez, Spark, Impala, Kafka, YARN, Zookeeper, Flume, Nifi |
| AWS Tools | IAM,S3,LAMDA,ATHENA ,Redshift,Glue,Glacier,CloudFormation, CloudWatch |
| Programming | Python, Scala, Unix, Java, YAML, BTEQ, PL/SQL |
| ETL Tools | IBM Data Stage, Informatica, SSIS |
| Reporting Tools | Tableau, OBIEE, MicroStrategy, ThoughtSpot, Alteryx |
| CICD/Version Tools | GIT, Bitbucket, Bamboo, Jenkins |
| Databases | Teradata, Oracle, SQL Server, Greenplum, Maria DB |
| No SQL Database | HBase, MongoDB |
| Scheduling Tools | Control-M, Autosys, Crontab |
| Domains | Finance, Retail, Insurance |
| Agile Tools | JIRA, Remedy, ServiceNow |
| Developer Tools | Win SCP, Putty, SQL Server 2014 Management Studio, Developer, Teradata Developer Studio, Greenplum developer tool |
| Operating Systems | Windows, LINUX, UNIX |

**Work Experience:**

**Bank of America, Addison, TX Oct 2021 to Till Date**

**Project: Client Writing & Client Underwriting Role: Cloud Data Engineer**

**Responsibilities:**

* Work with the business users, understand their business needs, analyze, and determine the feasibility of design within time and cost constraints. Evaluating various big data technology solutions to identify best fit for the current requirement.
* Analyze the existing system and connect with the infrastructure team to provide the Hadoop, AWS platforms along with AD groups, Databases in different environments like DEV, UAT and PROD.
* Designed data pipeline by usingPySpark and AWS tools like IAM, LAMBDA, Glue, SNS, CloudWatch and Redshift.
* Developing Spark applications using Spark-SQL in Databricks for data extraction, transformation and aggregation from multifile file formats for Analyzing and transforming the data to uncover insights into the cluster usage patterns.
* Designing, Building, installing, configuring, and supporting the Hadoop Cloudera and configure the same with Bit Bucket and Jenkins for CICD process.
* Migrating the CDH6.X environment to CDP7.X for all HDFS files, Hive queries and Spark jobs.
* Created architecture framework to process data from different sources to HDFS, HIVE, HBase and processed aggregated data to data mart for reports and predictive analytics using SQOOP, Kafka, Spark Core, Spark SQL, Scala, Hive, HBase, HUE and Impala.
* Created end-to-end data pipeline in cloud AWS by using AWS tools like S3, Lambda, Athena, AWS Glue, Redshift
* Created and stored the data into S3 buckets and designed Lambda function to trigger the Lambda script when the data loaded into S3 buckets
* Used Athena for analyzing the data in S3 buckets as well as AWS clue data catalog
* Created and scheduled the Glue jobs by creating the data catalog, crawl and applying the transformations on them for end-to-end data flow.
* Created tables and views in Redshift database and applied the optimization techniques in Redshift tables.
* Written complex UDFs using Python to process business logics and leverage UDFs in PySpark for computing large amounts of data.
* Push the code into Bitbucket once the development, unit testing and review completed and merge, build the repositories by using Jenkins to push the code into UAT
* Assisted the team with several coordination activities, actively participated in the daily stand ups, Planning, Reviews and Retrospectives in the two-week Iterations. And guided the team with several demos and knowledge sharing work sessions.

***Environment:*** Hadoop Cloudera(CDH5.X,CDH6.X & CDP7X),Databricks, Snowflake, HDFS, SQOOP, Spark, Spark SQL, S3, Glue, Athena, Redshift, Python, Hive, HUE, Impala, JIRA, SSIS, SQL Server, Oracle, Linux, GIT, Bit Bucket, Jenkins

**Wells Fargo, Dallas, TX Mar 2021 to Oct 2021**

**Project: RFDS Role: Big Data Engineer**

**Responsibilities:**

* Work with the business users, understand their business needs, analyze, and determine the feasibility of design within time and cost constraints.
* Analyzed the source system and raised the request for creating the AD groups if the source system is not existing.
* Raised the request for service id to access the AD groups if the AD groups are created newly.
* Sent mail to devOps team to create the git repo if the repo is new
* Checked git repo whether the repo has Flowmaster framework or not.
* Check all env folders are matched with Upstream and downstream connections
* Analysed the mapping sheet and create the json files which are input to the flowmaster framework.
* Analyzed the workflows are powered by offline batch jobs written in Scalding, a MapReduce-based framework
* Developed the Spark jobs by enhance scalability and performance, migrating the scalding and mapReduce jobs.
* Created hive external tables to load the data into HDFS files
* Created MongoDB collections and views
* Tested for both History and Go forward loads
* Worked and designed performance tuning of Spark applications for adjusting Batch Interval time, level of parallelism for spark jobs and memory tuning to get an optimized throughput for a longer running applications to mitigate expenses caused by inefficient use of resources.
* Created auto generated HQL scripts for Oracle to Hive column mapping, HIVE DDL scripts
* Push the code into git once the development, unit testing and review completed and merge, build the repositories by using Jenkins to push the code into UAT
* Created the implementation document for PROD release and same review with Support team
* Communicated with Support team and DevOps team on PROD release day to deploy the code into PROD.
* Run the post validation scripts and same confirmed by the business team.

***Environment:*** Hadoop MapR, HDFS, Flowmaster Framework, Spark Core, Spark SQL, Scala, Hive, HBase, HUE, JIRA, SQL Server, Oracle, Linux, GIT, Jenkins

**Bank of America, Addison, TX Aug 2019 to Mar 2021**

**Project: Client Writing & Client Underwriting Role: Big Data Engineer**

**Responsibilities:**

* Work with the business users, understand their business needs, analyze, and determine the feasibility of design within time and cost constraints. Evaluating various big data technology solutions to identify best fit for the current requirement.
* Analyze the existing system and connect with the Infrastructure team to provide the Hadoop, AWS platforms along with AD groups, Databases in different environments like DEV, UAT and PROD.
* Designing, Building, installing, configuring, and supporting the Hadoop Cloudera and configure the same with Bit Bucket and Jenkins for CICD process.
* Created architecture framework to process data from different sources to HDFS, HIVE, HBase and processed aggregated data to data mart for reports and predictive analytics using SQOOP, Kafka, Spark Core, Spark SQL, Scala, Hive, HBase, HUE and Impala.
* Created streams using Spark framework and processed real time data into RDDs and created analytics using SPARK SQL and Data frame API’s by using Python program to load data into HDFS.
* Analyzed existing ETL SSIS packages and prepared high and low-level design to create and migrate data into data lake by using Spark Core and Spark SQL with Python programming language.
* Written complex UDFs using Python to process business logics and leverage UDFs in pySpark for computing large amount of data.
* Worked and designed performance tuning of Spark applications for adjusting Batch Interval time, level of parallelism for spark jobs and memory tuning to get an optimized throughput for a longer running applications to mitigate expenses caused by inefficient use of resources.
* Created auto generated HQL scripts for Oracle to Hive column mapping, HIVE DDL scripts which are equal to Oracle tables, Custom scripts and DML scripts.
* Created views to populate the updated data to end business users.
* Pushed the Hive views to Impala by using invalidate or refresh them
* Push the code into Bitbucket once the development, unit testing and review completed and merge, build the repositories by using Jenkins to push the code into UAT
* Assisted the team with several coordination activities, actively participated in the daily stand ups, Planning, Reviews and Retrospectives in the two-week Iterations. And guided the team with several demos and knowledge sharing work sessions.

***Environment:*** Hadoop Cloudera, HDFS, SQOOP, Spark Core, Spark SQL, Python, Hive, HBase, HUE, Impala, JIRA, Rally, SSIS, SQL Server, Oracle, Linux, GIT, Bit Bucket, Jenkins

**Mphasis, USA APR 2019 to AUG 2019**

**Project: Historical Access Archive Role: Cloud Big Data Engineer**

**Responsibilities:**

* Work with the business users, understand their business needs, analyze, and determine the feasibility of design within time and cost constraints. Evaluating various big data technology solutions to identify best fit for the current requirement.
* Designed the pySpark program to read the data from different sources and applying complex transformations and then, load the data into AWS S3.
* Applied the conditions to archive the S3 data into Glacier
* Developed the LAMBDA and Glue ETL jobs to load the S3 data into Redshift
* Created the CloudFormation temple to create the Infrastructure to AWS Data Pipeline.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using CloudWatch.
* Created tables, views and copy the data from S3 in Redshift tables.
* Created the stack template by using CloudFormation tool
* Created AWS Data Pipelines by using IAM, EC2, EMR,S3, Glacier, LAMBDA, GLUE, ATHENA, Redshift in AWS tools.
* Scheduled the Hadoop jobs in Autosys and AWS jobs in Airflow

***Environment:*** Spark Core, Spark SQL, Python, Hive, HUE, IAM, EC2, EMR, S3, Glacier, LAMBDA, GLUE, ATHENA, Redshift, CloudFormation, Data Pipeline, JIRA, Oracle, Linux, Windows

**Anthem, Mason, OH APR 2018 to APR 2019**

**Project: RX Finance Role: Cloud Big Data Engineer**

**Responsibilities:**

* Providing directions on solutions, recommend options, drive things from design to development in big data/Hadoop.
* Process the claims data by applying the transformations in pyspark and provide the fixed length file to the end user.
* Created the pySpark programs to read the fixed length files and Oracle tables data by applying complex transformations on data frames and load the data into HDFS fixed length file for downstream process.
* Designed and implemented UDF's using Python for evaluation, filtering, loading, and storing of data in Spark data frames.
* Worked and designed performance tuning of Spark applications for adjusting Batch Interval time, level of parallelism for spark jobs and memory tuning to get an optimized throughput for a longer running applications to mitigate expenses caused by inefficient use of resources.
* Created Hive tables to capture the missing and invalid data from the source files.
* Push the code into Bitbucket and created the snapshot in the Bamboo.
* Load the on-premises data into AWS S3 by using Spark program.
* Run the Spark jobs by configure the EMR cluster
* Created the Data Pipeline to move the data landing S3 bucket to perm S3 bucket by using AWS tools like LAMBDA, Glue, ATHENA, Redshift
* Scheduled and Monitored the jobs in Control-M.

***Environment:*** Hadoop Cloudera, Spark, Python, HDFS, Hive, HUE, Impala, IAM, EC2, EMR, S3, LAMBDA, ATHENA, CloudFormation, JIRA, Oracle, Linux, GIT, Bit Bucket, Bamboo, UNIX, Windows

**Infosys Limited, USA FEB 2017 to APR 2018**

**Project: Global Data Portal Role: Technology Lead**

**Responsibilities:**

* Providing directions on solutions, recommend options, drive things from design to development in Big data/Hadoop.
* Designed the Spark Streaming application by using Scala Program for Producers and Consumers to push and pull the data from Kafka Topics
* Building data applications and automating the pipelines in Spark for bulk loads as well as Incremental Loads of various Datasets.
* Inserted the tables (raw data) into Hive for querying the large data systems and for result storage.
* Worked and designed performance tuning of Spark applications for adjusting Batch Interval time, level of parallelism for spark jobs and memory tuning to get an optimized throughput for a longer running applications to mitigate expenses caused by inefficient use of resources.
* Supervise complex data workflows based on data integration between Kafka, Apache Spark, HBase, Hive and other similar systems.
* Captured the data from Node.js jobs by using REST API for streaming data from Producers in KAFKA
* Reporting the spark job stats, monitoring, and running data quality checks are made available for each Datasets.
* Designed and developed scalable data pipelines using Apache Spark and Scala to process terabytes of data.
* Monitoring Spark containerized applications which are running on Jenkins using IBM monitoring Grafana.
* Improve the performance aspects of the spark applications by monitoring the CPU and memory utilization of the Jenkins over Grafana.
* Experienced in managing applications by executing and building them on the Maven Build tool and IDEs Eclipse and IntelliJ with a lot of supported plugins.
* Helped DevOps Engineers for deploying code and debug issues.

***Environment:*** Hadoop Hartonworks, HDFS, Hive, Spark Core, Spark SQL, Spark Streaming, Kafka, YARN, GIT, Bit Bucket, Jenkins, Bitbucket, Grafana, MongoDB, JIRA, UNIX, Windows

**Infosys Limited, Hyderabad, IND JAN 2015 to JAN 2017**

**Project: Data Café – Tech Foundation Client: Walmart, Bentonville, AR**

**Role: Technology Lead**

**Responsibilities:**

* Interacted with the business users for understanding the business requirements and the existing system and conducted the gap analysis to identify the feasibility of the existing system to accommodate the new integrations.
* Migrate Java code to Python script to standardize the database in future purpose and migrated bash scripts to point to SQL Server instead of Greenplum Database.
* Moved all SAP and NONSAP data flat files generated from various retailers to HDFS for further processing for historical load and daily loads.
* Written the Apache PIG scripts to process the HDFS data based on required output HDFS data.
* Created Hive tables to store the processed results in a tabular format.
* Developed the SQOOP scripts to make the interaction between Pig and Teradata, SQL Server Database.
* Created Hive tables to store the processed results in a tabular format.
* Developed the SQOOP scripts to make the interaction between Teradata, SQL Server Database.
* Created HBase tables to store the transactional data.
* Developed the UNIX shell scripts for creating the reports from Hive data.
* Setting up corn job to delete Hadoop logs/local old job files/cluster temp files.
* Scheduled these jobs in Crontab and Atomic scheduler to run the jobs daily or Weekly or Monthly.

***Environment:*** Hadoop Hartonworks, HDFS, Hive, HBase, SQOOP, UNIX, BTEQ, TPT, YAML, Teradata, SQL Server, Greenplum, Crontab, Autosys, GIT, Bit Bucket, Jenkins, JIRA, UNIX, Windows

**Infosys Limited, Hyderabad, IND SEP 2013 to JAN 2015**

**Project: Store Support & Enhancements Client: Starbucks, Seattle, WA, USA**

**Role: Data Engineer**

**Responsibilities:**

* Extensively involved in preparation of low-level design documents and unit test case preparation.
* Design Data Stage jobs for the business requirements using ETL specifications extracting data from heterogeneous source systems, transform and load into the target.
* Used Data Stage Designer and Data Stage Director for creating and implementing jobs, schedule running the jobs.
* Designed and coded the ETL logic using Data Stage to enable initial load and incremental processing from oracle and SQL.
* Create script for import and export the data by using Teradata utilities with BTEQ, FastLoad, MultiLoad and FastExport.
* Schedule the designed jobs in Control-M scheduler.
* Developed multiple tab reports in Tableau Desktop connecting to live or extract based on Historical data or Transactional data that were published to internal team for review.
* Providing OBIEE and Micro strategy technical support to BIAPPS reporting.
* Writing Unix Scripts to auto generate the jobs execution status.

***Environment:*** Data stage 9.1, Informatica, Tableau, OBIEE, MicroStrategy, Oracle, Teradata, Control-M, Remedy, JIRA, UNIX, Windows

**Cognizant Technology Solutions, Hyd, IND JUL 2009 to AUG 2013**

**Project: BSC Dev Client: Blue Shield of California, USA**

**Role: ETL Developer**

**Role and Responsibilities:**

* Extensively involved in preparation of low-level design documents and unit test case preparation.
* Extensively analyzed the subscriber, groups information against claims processing.
* Design Data Stage jobs for the business requirements using ETL specifications extracting data from heterogeneous source systems, transform and load into the target.
* Used Data Stage Manager for exporting & importing metadata definitions, creating Metadata definitions and view and edit the contents of the repository.
* Designed and coded the ETL logic using Data Stage to enable initial load and incremental processing from Oracle.
* Extensively involved in using the various activities like job activity, routine activity, wait for file activity, sequencer activity, termination activity, exception handler activity etc.
* Create script for import and export the data by using Teradata utilities with BTEQ, FastLoad, MultiLoad and FastExport.
* Involved in creating Shell Scripts.

***Environment:*** Data stage 9.1, Oracle, Teradata, Facets, Remedy, Windows, UNIX

**IP Creations Ltd, Chennai, IND May 2008 to Jul 2009**

**Client: Daman National Health Insurance, Abu Dhabi, UAE**

**Role: Java Documentum Developer**

**Roles and Responsibilities:**

* Extensively involved in preparation of low-level design documents and unit test case preparation.
* Good exposure in Connectivity Architecture of DocBrokers/Docbases/Servers.
* Exposure in Object hierarchy in Documentum as well as in Workflows.
* Involved in creating Cabinets, Folders and Documentum ObjectTypes and corresponding Attributes for those documents.
* Well versed with Import/Export/Copy/Move/Delete/CheckOut/Checkin and CancelCheckOut operations.
* Designed and implemented applications that support the new features using Java, JSP, Documentum, DFC, HTML, Java Script, and XML.
* Involved in creating Virtual documents and thus developing components using DFC and DQL programming.
* Excellent communication and presentation skills as well as vibrant team player.

***Environment:*** Java, JSP, CSS, HTML, Java Script, DFC, DQL, WDK, Webtop, Oracle, Windows 7, UNIX