**Sarat Chandra Bigdata Cloud Engineer**

**Email:** [**Saratcbigdata@gmail.com**](mailto:Saratcbigadata@gmail.com) **Contact: 469 846 8345**

**Visa: Green Card**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**PROFESSIONAL SUMMARY:**

* 11+ years of experience in IT which includes Analysis, Design, Datastage, Development of Big Data using Hadoop, AWS, Python, data Lake, Scala, design and development of web applications using JAVA, Spring boot and data base and data warehousing development using My SQL, Oracle
* Experience in creating applications using Spark with Python
* Experienced in Apache Spark and developing data processing and analysis algorithms using Python.
* Strong working experience on 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, Oozie, Airflow,NiFi(ETL)
* Building data processing triggers for Amazon S3 using AWS Lambda functions with Python.
* Worked on the Spark SQL and Spark Streaming modules of Spark and used Scala and Python to write code for all Spark use cases
* Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, 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.
* 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.
* Expertise in preparing Interactive Data Visualization's using Tableau Software from different sources
* Strong experience in analyzing large amounts of data sets writing PySpark scripts and Hive queries.
* For the app developing project, I implemented applications with Scala along with Akka and Play framework and implemented Restful services in Spring.
* Running of Apache Hadoop, CDH and Map-R distros, dubbed Elastic MapReduce(EMR) on (EC2)
* Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.
* Implemented DWH solutions to ingest data and load to target tables and adequately skilled with deployment tools
* Experience in migrating databases to Snowflake multi-cluster warehouses and adequate understanding of Snowflake cloud technology and using Snowflake Clone.
* Experience with Snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source system which include loading nested JSON formatted data into the snowflake table.
* Experienced with cloud: Hadoop-on-Azure, AWS/EMR, Cloudera Manager (also direct-Hadoop-EC2 (non EMR))
* Extensively worked on AWS services like EC2, S3, EMR, FSx, Lambda, Cloud watch, RDS, Auto scaling, Cloud Formation, SQS, ECS, EFS, DynamoDB, Route53, Glue etc.
* Hands on experience in VPN Putty and WinSCP, CI/CD(Jenkins)
* Creating service based applications using python to move data across accounts assuming the role(as direct access is denied)
* Experience in Data load management, importing & exporting data using SQOOP&FLUME.
* Experience in analyzing data using Hive, Pig and custom MR programs in Java.
* Expertise in Waterfall and Agile - SCRUM methodologies.
* Experienced with code versioning and dependency management systems such as Git, SVT, and Maven.
* Experience in Testing and documenting software for client applications.
* Writing code to create single-threaded, multi-threaded or user interface event driven applications, either stand-alone and those which access servers or services.
* Good experience in using Data Modelling techniques to find the results based on SQL and PL/SQL queries.
* Experience working with different databases, such as Oracle, SQL Server, MySQL and writing stored procedures, functions, joins, and triggers for different Data Models.
* Great team player and quick learner with effective communication, motivation, and organizational skills combined with attention to details and business improvements.
* Experienced in handling different file formats like Text file, Avro data files, Sequence files, Xml and Json files.

**TECHNICAL SKILLS**:

|  |  |
| --- | --- |
| Big Data Technologies | HDFS, Hive, MapReduce, Pig, Sqoop, Flume, Oozie, Hadoop distribution, and Hbase, Spark, Spark Streaming, Yarn, Zookeeper, Kafka, ETL.(Nifi, Talend etc.) |
| Databases | MySQL, MS-SQL Server 20012/16, Oracle 10g/11g/12c |
| Scripting/Web Languages | Pyspark, CSS3, XML, SQL, Shell/Unix, Perl, Python, Spark, Scala |
| Databases | Cassandra, HBASE, mongoDB, Oracle, MS SQL, Teradata. |
| Operating Systems | Linux, Windows XP/7/8/10, Mac. |
| Software Life Cycle | SDLC, Waterfall and Agile models. |
| Utilities/Tools | Eclipse, Tomcat, NetBeans, JUnit, SQL, SVN, Log4j, SOAP UI, ANT, Maven, Alteryx, Visio, Jenkins, Jira, Intellij. |
| Data Visualization Tolls | Tableau, SSRS, Cloud Health. |
| Cloud Services | AWS(EC2, S3, EMR, RDS, Lambda, Cloudwatch, FSx, Auto scaling, Redshift,Cloud Formation, Glue etc.) , Azure Databricks,Data Factory, Synapse Analytics, GCP, InfoWorks,Big Query. |

PROFESSIONAL EXPERIENCE

**CVS Health, Dallas, TX January2022 – June2023**

**Bigdata Engineer**

* Involved in Requirement gathering, Business Analysis and translated business requirements into Technical design in Hadoop and Big Data.
* Developed Pyspark scripts to extract the data from the web server output files to load into GCS buckets within the GCP environment.
* Created various data pipelines using Spark, Pyspark and SQL for faster processing of data
* Designed number of partitions and replication factor for Kafka topics based on business requirements and worked on migration with Spark transformations using Spark and Scala, initially done using python (PySpark).
* Written Spark-SQL and embedded the SQL in Pyspark script to generate jar files for submission onto the Hadoop cluster and Worked with a team to migrate from Legacy/On prem environment into GCS, the storage within the GCP.
* Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format Parsed Semi Structured JSON data and converted to Parquet using Data Frames in PySpark.
* Developed a Python Script to load the CSV files into the (GCP/)GCS buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Worked on migrating the data warehouse to the the Adobe big query tables in the RAW layer which further are made available to the consumer teams via Big query view setup on the tables.
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions, Documented the systems processes and procedures for future references.
* Created Dockerized backend cloud applications with exposed Application Program Interface (API) interfaces and deployed on Kubernetes.
* Experienced in writing live Real-time Processing using Spark Streaming with Kafka
* Worked with querying data using SparkSQL on top of Spark engine.
* Involved in managing and monitoring Hadoop cluster using Cloudera Manager.
* Used Python and Shell scripting to build data pipelines.
* Involved in the migration strategy to move data from hive on-prem to the cloud storage within the GCP environment using the Replicator Infoworks tool
* Developed and executed a migration strategy to move Data Warehouse from Oracle platform to GCS, within the GCP environment using InfoWorks tool.
* Involved in making the data files available to the consumer teams via Big Query view setup on the Adobe Big Query tables in the raw layer using the Infoworks ETL tool
* Assisted in Cluster maintenance, cluster monitoring, adding and removing cluster nodes for data cleaning and pre-processing and TIDAL automation.
* Created data partitions on large data sets and DDL on partitioned data.

**Environment**: Scala, Datastage, Spark, Cloudera, SQL, RDBMS, Python, data Lake,Jira, Confluence, Shell scripting, NIFI, GCS/GCP, Git, Kafka, CI/CD(Jenkins), Kubernetes, Infoworks, Informatica, Big Query, Replicator Infoworks, Tidal automation.

**AT&T, Dallas, TX May 2020 – December2021**

**Bigdata Cloud Engineer**

* Involved in Requirement gathering, Business Analysis and translated business requirements into Technical design in Hadoop and Big Data.
* Involved in SQOOP implementation which helps in loading data from various RDBMS sources to Hadoop systems.
* Developed Python scripts to extract the data from the web server output files to load into HDFS.
* Involved in HBASE setup and storing data into HBASE, which will be used for further analysis.
* Worked on Written a python script which automates to launch the EMR cluster and configures the Hadoop applications using boto3.
* Created various data pipelines usingSpark, Scala and SparkSQL for faster processing of data
* Designed number of partitions and replication factor for Kafka topics based on business requirements and worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Implemented Spark using Python and Spark SQL for faster processing of data and Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark)
* Written Spark-SQL and embedded the SQL in SCALA files to generate jar files for submission onto the Hadoop cluster
* Created Data Quality Scripts using SQL and Hive to validate successful das ta load and quality of the data. Created various types of data visualizations using Python and Tableau.
* Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format Parsed Semi Structured JSON data and converted to Parquet using Data Frames in PySpark.
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions, Documented the systems processes and procedures for future references.
* Involved in Configuring Hadoop cluster and load balancing across the nodes.
* Involved in Hadoop installation, Commissioning, Decommissioning, Balancing, Troubleshooting, Monitoring and, debugging Configuration of multiple nodes using Hortonworks platform.
* Involved in working with Spark on top of Yarn/MRv2 for interactive and Batch Analysis
* Worked with a team to migrate from Legacy/On prem environment into AWS.
* Created Dockerized backend cloud applications with exposed Application Program Interface (API) interfaces and deployed on Kubernetes.
* Stored and retrieved data from data-warehouses using Amazon Redshift.
* Experienced in analyzing and Optimizing RDD's by controlling partitions for the given data
* Experienced in writing live Real-time Processing using Spark Streaming with Kafka
* Used HiveQL to analyze the partitioned and bucketed data and compute various metrics for reporting
* Worked with querying data using SparkSQL on top of Spark engine.
* Involved in managing and monitoring Hadoop cluster using Cloudera Manager.
* Used Python and Shell scripting to build pipelines.
* Developed data pipeline using SQOOP, HQL, Spark and Kafka to ingest Enterprise message delivery data into HDFS.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Pig and Hive.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Assisted in creating and maintaining Technical documentation to launching HADOOP Clusters and even for executing Hive queries and Pig Scripts.
* Assisted in Cluster maintenance, cluster monitoring, adding and removing cluster nodes and Installed and configured Hadoop, Map Reduce, HDFS, developed multiple Map Reduce jobs in java for data cleaning and pre-processing.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Automated and monitored complete AWS infrastructure with terraform.
* Created data partitions on large data sets in S3 and DDL on partitioned data.
* Converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size.

**Environment**: HDFS, Hive, Scala, Sqoop, Datastage, Spark,Tableau, Yarn, Cloudera, SQL,Terraform, Splunk, RDBMS, Python, Elastic search, data Lake, Kerberos, Jira, Confluence, Shell/Perl Scripting, Zookeeper, NIFI, AWS(EC2, S3, EMR, Redshift, ECS,Glue, S3, VPC, RDS etc.), Ranger, Git, Kafka, CI/CD(Jenkins), Kubernetes, Azure Databricks.

**Trimble, Dallas, TX March 2019 – April 2020**

**Bigdata AWS Cloud 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
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala
* 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.
* Troubleshooting data load issues.
* 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, Metastore 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, NIFI, Spark, Scala, Hive, Cloudera CDH4, Oracle, Elastic search, Kerberos, Datastage, SFTP, data Lake, Impala, Jira, Wiki, Alteryx, Teradata, Shell/Perl Scripting, Kafka, AWS EC2, S3, EMR, Cloudera.

**Alcons Lab, Fort Worth, TX  Dec 2017 – Feb 2019**

**Bigdata Engineer**

**Responsibilities:**

* Configured real-time streaming pipeline from DB2 to HDFS using Apache Kafka.
* Followed Agile Scrum methodology that included iterative application development, weekly Sprints and stand up meetings
* Acquire data from Transactional source systems to Redshift data warehouse using spark and AWS EMR
* Creating External and Managed Hive tables and working on them using HiveQL.
* Analysis &documenting of new/Existing system.
* Validated the Map reduce, Pig, Hive Scripts by pulling the data from the Hadoop and validating it with the data in the files and reports.
* Provided a new Web Service and Client using Spring-WS to get the alternate contractor details.
* Used JIRA for the issue tracking and bug reporting.
* Utilized several Java 8 concepts like Stream API, Time API, Collection API, lambda expressions to migrate the existing application.
* Expertized in implementing Spark using Scala and Spark SQL for faster testing and processing of data responsible to manage data from different sources.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data
* Collaborated with the team using GIT, GitHub, SourceTree version control platform.
* As part of another project I made use of Amazon Web Services (AWS).
* Bottle micro-framework implemented with REST API and MongoDB (NoSQL) as back end database.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Developed Spark code to using Scala and Spark -SQL for faster processing and testing.
* Designed and created Hibernate persistence classes using Hibernate API.
* Wrote Stored Procedures/Triggers/Functions using SQL Navigator to perform operations on Oracle database.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Automated all the jobs from pulling data from databases to loading data into SQL server using shell scripts.
* Installed Name node, Secondary name node, Yarn (Resource Manager, Node manager, Application master), Data node.
* Used Scala to convert Hive/SQL queries into RDD transformations in Apache Spark.
* Developed ETL process using Jitterbit Harmony cloud Integration tool.
* Good experience in troubleshooting production level issues in the cluster and its functionality.
* Implemented AWS solutions using EC2, S3, RDS, EBS, Elastic Load Balancer, Auto-scaling groups.
* Worked on Apache Flume to stream data from Oracle to Apache Kafka topics.
* Managed docker images using Quay.
* Installed Oozie workflow engine to run multiple Hive and Pig Jobs.
* Created hive managed and external tables.
* Exporting data to Teradata using SQOOP.
* Monitoring systems and services through Cloudera Manager to make the clusters available for the business.
* Developed SQOOP scripts to load data from Oracle to Hive external tables.

**Environment**:Hadoop, HDFS, Data Warehouse, Pig, Hive, Spark, Scala, MapReduce, Java, Cloudera Hadoop, Dell Boomi, HDFS, Map Reduce, Hive, Sqoop, SSIS, SQL, Oozie, Grafana.

**T. Rowe Price, Owing Mills, MD Jan 2016 – Nov 2017**

**data/Scala Developer**

**Responsibilities:**

* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala
* Developed Real time data processing applications by using Scala and Python and implemented Apache Spark Streaming from various streaming sources like Kafka, Flume and JMS
* Developed multiple POCs using Scala and deployed on the Yarn cluster, compared the performance of Spark, with Hive and SQL/Teradata.
* Analysed the SQL scripts and designed the solution to implement using Scala.
* Developed analytical component using Scala, Spark and Spark Stream.
* Developing UDFs in java for hive and pig and worked on reading multiple data formats on HDFS using Scala.
* Troubleshoot and debug Hadoop ecosystem run-time issues.
* Used Oozie Scheduler systems to automate the pipeline workflow and orchestrate the map reduce jobs that extract
* Worked with Hue GUI in scheduling jobs with ease and File browsing, Job browsing, Metastore management.
* Worked with BI team in the area of Big Data Hadoop cluster implementation and data integration in developing large-scale system software.
* Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, HBase, Flume, Oozie Zookeeper and Sqoop.
* Responsible for cluster maintenance, adding and removing cluster nodes, cluster monitoring and troubleshoot managing and reviewing data backups and Hadoop log files.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Extensively involved in Installation and configuration of Cloudera distribution Hadoop, NameNode, Secondary NameNode, JobTracker, TaskTrackers and DataNodes.
* Created POC to store Server Log data in MongoDB to identify System Alert Metrics.
* Monitored Hadoop cluster job performance, performed capacity planning and managed nodes on Hadoop cluster.
* Worked with application teams to install operating system, Hadoop updates, patches, version upgrades as required.
* Loaded data into the cluster from dynamically generated files using Flume and from relational database management systems using Sqoop.
* Performed analysis on the unused user navigation data by loading into HDFS and writing MapReduce jobs. The analysis provided inputs to the new APM front end developers and lucent team.
* Wrote MapReduce jobs using Java API and Pig Latin.
* Wrote Pig scripts to run ETL jobs on the data in HDFS and further do testing.
* Used Hive to do analysis on the data and identify different correlations.
* Involved in HDFS maintenance and administering it through Hadoop-Java API.
* Written Hive queries for data analysis to meet the business requirements.
* Automated all the jobs, for pulling data from FTP server to load data into Hive tables, using Oozie workflows.
* Involved in creating Hive tables & working on them using HiveQL and perform data analysis using Hive and Pig.
* Used Qlikview and D3 for visualization of query required by BI team.
* Defined UDFs using PIG and Hive in order to capture customer behavior.
* Design and implement MapReduce jobs to support distributed processing using java, Hive and Apache Pig.
* Create Hive external tables on the MapReduce output before partitioning, bucketing is applied on it.
* Loaded the load ready files from mainframes to Hadoop and files were converted to ASCII format.
* Configured Hive Server (HS2) to enable analytical tools like Tableau, Qlikview and SAS to interact with Hive tables**.**

**Environment:** Hadoop, MapReduce, HDFS, Hive, Java, SQL, Cloudera Manager, Pig, Sqoop, ZooKeeper, Teradata, PL/SQL, MySQL, Hbase, Datastage, ETL(Informatica/SSIS).

**TMW Systems, Dallas,TX Sept 2011 - Dec 2015**

**Java SQL Developer**

**Responsibilities:**

* Involved in all the phases of SDLC process. Involved in creating software requirement specification document
* Developing interfaces that integrates billing with Simulators using EJB stateless session beans.
* Used EJB Entity Beans to maps entity objects such as customer, account and product to relational database table i.e. Oracle.
* Design and development of user Interfaces using JSP, HTML, CSS, JavaScript, AJAX.
* Deploying wars on Weblogic application server and granting access to users.
* Create RFP (Request for Proposal) Microservice to provide RESTful API utilizing Spring Boot with Spring MVC.
* Designed and Developed Microservices (By breaking down monolith WebServices) using Spring Boot.
* Implemented Struts framework (Action & Controller classes) for dispatching request to appropriate classes.
* Used simple Struts Validation for validation of user input as per the business logic and initial data loading.
* Analyzed the data and system requirements, conducted meeting with QA team for writing test conditions and test scripts.
* Developing SOAP web services to be shared with CRM to interact with Prime Biller
* Creating the DEV build and resolving different build issues.
* Ran check styles, PMD Defects, find bugs etc and fixed them if there are any.
* Used XML and XSL extensively as the script logic was completely separated from the UI.
* Configured EMMA tool and ran test suite and made sure that 100% test coverage.
* Implemented Maven as build and configuration tool.
* Develops microservices and has extensive experience using GitLab, Jenkins, Clustering other tools and technologies for developing a scalable application.
* Co-coordinating with QA team during the QA phase of implementation.

**Environment:**Java, Servlets, JSP, EJBs, JavaScript, CRM, AJAX, SOAP, Struts, Web logic, Oracle-SQL, P/LSQL, TOAD, Microservices, Eclipse, HTML, UNIX.

**Education Details:**

Bachelor of Engineering, Electrical and Electronics Engineering, (2002-2006), Osmania University Hyderabad, India