# Senior Data Engineer

**Name: Chitra Lekha**

**Contact: +1 862-294-4997**

**Email ID:** lekhasrinivas12@gmail.com

## PROFESSIONAL PROFILE

- 10+ years of professional IT experience in BIG DATA using HADOOP framework and Analysis, Design, Development, Documentation, Deployment, and Integration using SQL and Big Data technologies as well as Java / J2EE technologies with AWS, AZURE
- Strong experience developing Spark applications using Spark SQL, PySpark, and Delta Lake in Databricks for data extraction, transformation, and loading from multiple file formats for visualization and analysis.
- Competency in designing pipelines to extract data from various sources, transforming data according to analytics requirements using Data Flow, and loading refined data to desired destinations in Azure Data Factory (ADF).
- Expert in performing incremental loading in Data Factory using control table and watermarks.
- Experienced in provisioning Azure Databricks clusters and installing required packages and libraries.
- Proficient in working with Delta Lake (Delta format) and Parquet files in Azure Databricks with the Delta Lake: ingesting data from raw sources to Bronze tables, cleaning and transforming data into Silver tables, and creating Gold tables tailored to specific business needs.
- Hands-on experience with AWS (Amazon Web Services), Elastic Map Reduce (EMR), RedShift, Amazon RDS, Glue, Storage S3, EC2 instances GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Data Proc, Stack driver, and Data Warehousing.
- Design and implement DevOps tools using CI/CD architecture and automation solutions.
- Experience in using SSIS tools like Import and export wizard, package installation and SSIS package designer.
- Designed CI/CD pipeline with docker and GitHub by virtualizing servers through automation.
- Build and release DevOps in automating, deploying, building and releasing code in various environments.
- Expertise in server side development for UI using JSP and Angular.
- Experienced in building and maintaining pipelines for deriving data in Azure Data Factory from OLTP to ODS, from ODS to Integration, and from Integration to OLAP.
- Comprehensive experience in UNIX shell scripting, administration work in UNIX/ LINUX(in Exadata ) and DB2.
- Expertise in T-SQL (DDL, DML, TCL), and developing SQL Server Programmability Objects such as Stored Procedures, Functions, Triggers, Views, and Sub queries for various business requirements.
- Familiar with dimensional modelling and strong hands-on experience with Star Schema and Snow-Flake Schema for the fact, and dimension tables in Data Warehouse.

- Experience in creating configuration files to deploy SSIS packages across all environments.Transformed data from one server to another using tools like bulky copy programs and SSIS (2005/2008).
- Vast expertise in setting up, maintaining, and administering Amazon Web Services (AWS) features like EC2, S3, Redshift, Glue, and Athena processing High availability, fault tolerance, Scalability, Elastic Beanstalk (EBS), Auto Scaling, Security Groups, Dynamo DB, Red shift, CloudWatch, CloudFormation, CloudTrail, CloudFormation, Ops Works, Kinesis, IAM, SQS, SNS, SES, and Code Commit, Code Pipeline, CloudFormation, Code Build, Code Deploy.
- Skilled in T-SQL query performance optimization under SQL Server Management Studio using Tuning Advisor, Execution Plan, Trace Flags, and Extended Events.
- Proficient use of Python, Scala, and Java to create Spark applications for interactive analysis, batch processing, and stream processing, understanding of using MapReduce applications and Avro tools to process Avro data files.
- Knowledge of backend programming oracle (PL/SQL), database design, data analysis, data modelling, data migration, data refresh, and performance tuning.
- Knowledge in developing applications in single page using javascript frameworks like AngularJS and CSS.
- Experience in SAS reports and exporting it to excel spreadsheets and create web reports using web report studio.
- Built scalable docker infrastructure for micro services using AWS ECS with fargate.
- Has used Kafka and Kafka brokers to start spark context and process livestreaming.
- Expertise in creating reports and dashboards in Tableau and knowledge of data mining and warehousing using ETL tools (BI Tool).
- Proficient use of several ETL technologies, such as Informatica Power Centre ELT tool, for data migration, data profiling, ingestion, data cleaning, transformation, import, and export.
- An in-depth comprehension and familiarity with NoSQL databases including MongoDB, PostgreSQL, HBase, cloud-native distributed SQL database CockroachDB, and Cassandra.
- Extensive knowledge in the fundamentals of Java, Scala, SQL, PL/SQL, and Restful web services.
- Expertise in creating unique UDFs for Pig and Hive that incorporate Python/Java functionality into Pig Latin and HQL (HiveQL), as well as experience using UDFs from the Piggybank UDF Repository.
- Worked on various reporting tools like Power BI, Tableau & Qlik Sense to showcase the insights of the data to stakeholders.
- Expertise in working with Terraform key features such as Infrastructure as code , execution plans, resource graphs, change automation and extensively used Auto scaling launch configuration templates for launching EC2 instances while deploying microservices.
- Great understanding of SDLC methodologies like Agile and Waterfall and worked in Agile Scrum environments.
- In depth understanding of windows Powershell scripting.

- Created complex mappings in Talend using components like tMap, tJoin, tReplicate, aggregate, tBugger.
- Hands on experience on Talend troubleshooting and database to understand to understand the errors in job and used the tMap and expressions editor to evaluate complex expressions and look at the transformed data to solve mapping issues.

**Technical Skills:**

| | |
|---|---|
| **Hadoop Components / Big Data** | Hadoop, HDFS, MapReduce, PIG, Hive, HBase, Sqoop, Impala, Flume, Kafka, Yarn, Cloudera Manager, PySpark, Airflow, Kafka, Snowflake . |
| **Languages** | Scala, Python, SQL, JSP, Angular,Shell Scripting, Hive QL, UNIX , LINUX. |
| **BI Tools** | SSIS, SSRS, SSAS. |
| **Cloud platform** | AWS (Amazon Web Services), Microsoft Azure , GCP |
| **Reporting & ETL & Other Tools** | Tableau, Talend, AWS GLUE, Pentaho, Informatica Power ELT, SAP Business Objects, Web Intelligence. |
| **Databases** | Oracle, SQL Server, MySQL, MS Access, NoSQL Database ( HBase, Cassandra, Mongo DB), Erwin, Visio, CockroachDB, Poster SQL, DB2,Oracle 12c/11g/10g, Oracle Exadata. |
| **Big Data Technologies** | Hadoop, HDFS, Hive, Pig, Oozie, Sqoop, Spark, Teradata SQL, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Flume, Airflow, Informatica ELT, Snowflake, Data Bricks, Kafka, Cloudera, Presto, Terraform |
| **Containerization** | Kubernetes, TOAD, MS Office, BTEQ, Teradata SQL Assistant. |
| **CI & Reporting Tools** | Jenkins, Business Objects, Crystal Reports |
| **Operating Systems** | UNIX, LINUX, Windows, MacOS. |
| **Other Tools** | TOAD,CI/CD,SQL PLUS, Oracle PL/SQL, MS Visio, Bitbucket, MS Office, SQL LOADER, Windows Powershell scripting, Apache airflow, meson, etc. |

## Professional Experience:

**Client: Barclays, Whippany, NJ. Oct 2021 – Till Date**
**Role: Senior Data Engineer**

**Responsibilities:**

- Took a major part in analyzing business requirements and preparing comprehensive specifications by project guidelines needed for project development.
- Disk and file system management through Solstice disk suite on Solaris and another logical volume manager for other flavor of UNIX.
- Designed and constructed an ETL framework to load data into Hive from MongoDB and MySQL, and back into MongoDB from Hive.
- Used Talend to trigger jobs to AWS for snowflake external staging table ingestion.
- Used Talend to monitor jobs and schedule it for snowflake ETL operations.
- Design maintain and implement data pipeline tooling, developing best practices for data collection, ingestion, transformation and storage as part of data Operation functions(DataOps)
- Worked extensively on MongoDB, Hadoop-Hive, Spark, SQLs, and PySpark.
- Created Spark applications for data quality assessment using Scala and Java by using higher-order functions.
- Developed new or modified existing SAS programs to load data from the source and create study specific dataset.
- Utilized SAS procedures, macros and other SAS applications for data extraction, data cleansing and reporting.
- Done T-SQL query performance optimization under SQL Server Management Studio using Tuning Advisor, Execution Plan, Trace Flags, and Extended Events.
- Creating pipelines, data flows and complex data transformations and manipulations using PySpark with Databricks(DataOps)
- Used ETL(SSIS) to develop jobs for extracting, cleaning, transforming and loading data into data warehouse.
- Helped team to setup their repositories in bit bucket and set up jobs to help make use of CI/CD environment.
- Setup full CI/CD pipelines so that each commit a developer makes go through standard process of software lifecycle and gets tested well enough before it can make it to the production(DataOps).

- Developed spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing & transforming the data to uncover insights into the customer usage patterns.
- Responsible for estimating the cluster size, monitoring and troubleshooting of Spark Databricks cluster.
- Utilizing load balancers for auto-scaling network load, EC2 for providing compute resources, S3 buckets to store data in compressed object form partitioned by keys, CloudFormation for provisioning of data with yml file and creating AWS credentials.
- Configured the storage on S3 buckets and installed the application on AWS EC2 instances with allocation of AMI, selection of instance types, storage (EBS, EFS) and configuration of instance using security groups, key pairs, IAM roles, placement groups.
- Used Pentaho import export utility to migrate Pentaho Transformations and job from one environment to others.
- Used Pentaho design studio for creating custom parameters as well as generating reports for visualization.
- Used Pentaho report designer to create various reports having drill down functionality by creating reports and drill through functionality by creating sub-reports within the main reports for data visualization.
- Prepared the complete data mapping for all the migrated jobs using SSIS.
- Involved in designing, developing and deploying reports in MS SQL Server environment using SSRS-2008 and SSIS in business intelligence development studio.
- Installed and configured Pentaho BI suite 4.2 & 4.4 along with Enterprise Repository in Pentaho BI server.
- Experience with Snowflake cloud data warehouse ELT and AWS S3 bucket for integrating data from multiple source systems, including loading nested JSON.Implemented microservices using Spring boot, spring based microservices, and enabled discovery using Netflix eureka server
- Developed Spark applications in Python (PySpark) on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
- Followed the organization defined naming convention for naming flat file structure, talend jobs, and daily batches for executing Talend jobs.
- Experience in cloud data migration using AWS and Snowflake for ELT operations. And strong experience in migrating other databases to Snowflake for extraction, loading and transformations(ELT).
- Act as subject matter expert on Dev Ops best practices with CloudFormation auto scaling groups.
- Developed AWS Lambda to spin up and teardown AWS EMR cluster using Service Catalog and implement Auto Scaling by using scaling policies for cluster efficiency.
- Worked on transforming Hive/SQL queries into Spark transformations using Python, Scala, and Spark RDDs
- Using the Apache Airflow ELT tool on GCP, developed Python code for each job's SLA observer, time sensor, dependencies, and various workflow management and automation activities.

- Performed Data cleaning followed by Exploratory Analysis through visualization on real-time datasets.
- Used various techniques to build a model that could improve its accuracy in generating better predictions.
- 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 PySpark scripts to convert huge volume files to parquet format and Performed Data Analysis Using PySpark (SparkSQL).
- Authored several scripts leveraging VMware's PowerCLI and Windows Powershell for capacity planning and monitoring of virtualization infrastructure.
- Automated Microsoft recommended load balancing techniques using powershell to reduce workload .
- Imported the data from various formats like JSON, Sequential, Text, CSV, AVRO and Parquet to HDFS cluster and produced them with compression for optimization using PySpark.
- Used PySpark and Developed Scripts to Filter & calculate aggregate Data and write the resulting data to a variety of file formats with compression for optimization as per business needs.
- Involved in developing real time streaming applications using Pyspark , HIVE and Apache Flink on Hadoop cluster.
- Deployed the project on Amazon EMR with S3 connectivity for setting backup storage.
- Developed Pyspark scripts utilizing SQL and RDD in spark for data analysis and storing back into S3.
- Developed code in Spark SQL for implementing Business logic with python as programming language.
- Designed, Developed and Delivered the jobs and transformations over the data to enrich the data and progressively elevate for consuming in the Pub layer of the data lake.
- Worked on Sequence files, Map side joins, bucketing, partitioning for hive performance enhancement and storage improvement.
- Wrote, compiled, and executed programs as necessary using Apache Spark in Scala to perform ETL jobs with ingested data.
- Built scalable docker infrastructure for micro services utilizing ECS with fargate by creating task definition JSON files.
- Implemented cluster services using Kubernetes and docker to manage local deployment in Kubernetes by building a self hosted Kubernetes cluster using Terraform and deploying application containers.
- Designed solutions to process high volume data stream ingestion, processing low latency data provisioning using Hadoop ecosystem Spark, Scala, Druid.
- Maintained Kubernetes patches and upgrades. Managed multiple Kubernetes clusters in a production environment.
- Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation and used Spark engine, Spark SQL for data analysis and provided it to the data scientists for further analysis.
- Data Integrity checks have been handled using hive queries, Hadoop, and Spark.
- Worked on performing transformations & actions on RDDs and Spark Streaming data with Scala.

- Implemented the Machine learning algorithms using Spark with Python.
- Designs and implements Scala programs using Spark Data frames and RDDs for transformations and actions on input data.
- Improved the Hive queries performance by implementing partitioning and clustering and Optimized file formats (ORC).
- Worked with Spark to improve performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD, Spark YARN.
- Experienced in writing machine learning live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
- Developed ETL process in AWS glue by defining catalog, writing crawlers to migrate data from external sources like S3 , Text/Parquet Files into AWS redshift for storage and scheduled jobs using AWS event bridge scheduler.

**Environment**: ETL framework, Informatica, Angular, CI/ CD, JSP, Apache Spark, MongoDB, Airflow ELT tool, MySQL, Spark RDDs, PySpark, Kafka, ECS fargate, Flink, DynamoDB, Auto Scaling, Spark YARN, Lambda, Nifi,Druid, Snowflake, Machine learning Python, AWS Aurora Apache Kafka, Spring framework, Java,Matillion, Databricks, Neo4J, Aurora, Fivetran, Cloudformation, SAS, CockroachDB, windows powershell scripting, SSIS, terraform, Data Ops, Pentaho, Vertica, GCP

**Client: Smith Micro, Pittsburgh, PA.** **Mar 2019 – Sep 2021**
**Role: Principle Data Engineer**

**Responsibilities:**

- Implemented Spark Scala applications for data transformations and to optimize query execution.
- Develop utilities using scala and Python to minimize number of lines of code in daily developments.
- Working on experimental Spark API for better optimization of existing algorithms such as Spark context, Spark SQL, Spark Streaming, Spark Data Frames.
- Responsible for developing Python wrapper scripts that will extract specific date range usingSqoop by passing custom properties required for the workflow
- Optimize the Pyspark jobs to run on Kubernetes Cluster for faster data processing.
- Develop Python, PySpark, and Spark scripts to filter/cleanse/map/aggregate data and data ingestion.
- Implemented informatica recommendations, methodologies and best practices.
- Used transformations like filter, expression, joiner, normalizer, sorter and union to develop robust mappings in the Informatica designer.
- Created Talend development standards to describe the guidelines for Talend developers and team members for naming conventions to be used in the transformation and also development and production environment structures

- Responsible for developing, support, and maintenance of ETL(Extract, Transform, Load) processes using talend Integration Suite.
- Created SSIS package for file transfer from one location to another using FTP task.
- Developed mappings in Informatica using BAPI and ABAP function calls in SAP.
- Develop various analytical queries on top of Hive database to generate reports for business needs.
- Worked in migration of RDMS data into Data Lake applications.
- Implemented Optimization techniques while using Hive like Partitioning and Bucketing etc.
- Develops applications to import and export between cloud services like Amazon S3.
- Developed microservices to consume data from REST end point and load into Kafka topic.
- Experience in using Pentaho report designer and performing data masking/ protection using Pentaho Data Integration.
- Used Pentaho data Integration to design all ETL processes to extract data from various sources including live system and external files, cleanse and then load the data into the target data warehouse.
- Developed Spark Streaming application to consume from Kafka topic and write into Hive table.
- Automated data pipelines using scheduler like Apache Oozie.
- Created SSIS package for loading the data coming from various interfaces like Orders, Adjustments, and objectives and also used multiple transformations in SSIS to collect data from various sources.
- Worked on connecting Cassandra database to the Amazon EMR File System for storing the database in S3.
- Worked on AWS cloud S3, RDS, Athena, Cloud Watch, EC2, IAM policies, SQS, Lambda, CloudFormation ,AWSSage maker
- Worked on Apache NiFi and Druid for executing Spark script, Sqoop scripts through NiFi, worked on creatingscatter and gather pattern in NiFi, Ingesting data from Postgres to HDFS, Fetching Hive metadataand storing in HDFS, created a custom NiFi processor for filtering text from Flow files.
- Design, develop, and test dimensional data models using Star and Snowflake schema methodologies under the Kimball method.
- Created Tableau dashboards/reports for data visualization, Reporting and Analysis and presented it to Business.
- Worked extensively with DATA MODELING, DATA MIGRATION, DATA CLEANSING, DATA PROFILING, and ETL Process features for data warehouses.
- Implemented a server less architecture using API Gateway to expose the services, Lambda functions to make the connections between the services, and Dynamo DB as a database repository for storage and deployed AWS Lambda code from Amazon S3 buckets with IAM roles for running the application.
- Involved in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, SQOOP, Apache Spark, with Cloudera Distribution
- Written automation scripts for creating resources in Open stack cloud using python and terraform modules.

- Integrated Apache Airflow ELT with AWS to monitor multi-stage ML workflows with the tasks running on Amazon Sage Maker.
- Developed scripts with Windows Powershell to automatically configure network settings and vmkernal ports.
- Data type mappings between Vertica database and Oracle.
- Prepared Conceptual, logical and physical ER models using Erwin Data modeler.
- Analyze various type of raw file like Json, Csv, Xml with Python using Pandas, Numpy etc.
- Involved on configuration, development of Hadoop environment with AWS cloud such as EC2,EMR, Redshift, Route 53, Cloud watch.
- Experience in using CDC tools like IBM CDC for incremental update.
- Created CI/CD pipeline using DevOps tools like Jenkins, GitHub(Bitbucket) for Continuous integration and deployment for destination end points.
- Developed spark optimized data pipelines using scala and python.
- Expert in building bash, shell scripting, Python for various functionalities.
- Developed ingestion pipelines for pulling data from AWS S3 buckets to HDFS for further analytics.
- Developed Lambda functions to trigger ETL jobs across AWS tools.
- Responsible for implementing monitoring solutions in Ansible, Terraform, Docker, and Jenkins.
- Creating Athena glue tables on existing csv data using AWS crawlers.
- Also worked on L3 production support for existing products
- Practiced and evangelized agile development approaches.
- Worked on agile environment, used GitHub for version control and Teamcity for continuous build
- Evaluated Matillion and Fivetran for streaming and batch ingestion into Snowflake for ELT
- Used Jersey framework to implement JAX-RS ( Java API for RESTful service and XML)
- Used a light front-end framework against JSON API for their service request.
- Used Meson for workflow and scheduling tasks.

**Environment:** HDP, HDFS, Hive, Spark, Oozie, HBase, AWS, Scala, Python, Bash, Kafka, Java, Jenkins, Spark Streaming, Tez, AWS Athena, Glue,Matillion, Fivetran and DBT, Spring framework, Informatica,Vertica, Druid, DevOps, Windows Powershell scripting, SSIS, Meson, Apache airflow, terraform, Pentaho, Vertica, GCP, Talend

**Client: MasterCard, Houston, TX** **Sep 2016 – Mar 2019**

**Data Engineer**

**Responsibilities:**

- Assist in determining the requirements and functionality required for a project by working with the technical staff team, business managers, and practitioners in the business unit on Spark Data Frames, conducting wide and narrow transformations and operations like filter, lookup, join, count, etc.
- Utilized PySpark and Spark Streaming with Data Frames to work with Parquet files and ORC.

- For the needs of a functional pipeline, batch and streaming processing applications were created utilizing Spark APIs.
- Developed PySpark code that creates data frames from raw layers with Avro formatting using Spark SQL and writes them to internal tables in the data service layer in Parquet format.
- The creation of workflows using Apache Airflow ELT, followed by the scheduling of Hadoop tasks that govern huge data transformations using Apache Oozie.
- Proficient use of Sqoop to import and export data from relational databases and Teradata into HDFS/Hive
- Using Azure Databricks and Data Factory, create and manage an ideal data pipeline architecture on the Microsoft Azure cloud.
- Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse
- Demonstrated ability working and adapting to Big Data tools such as Databricks, blob storage, Data bricks, HDFS, Pig, MapReduce Hive.
- Built scalable docker infrastructure for micro services using AWS ECS with fargate.
- Created Python Spark scripts on Azure HDInsight for data aggregation, validation, and performance testing.
- Automated jobs using different triggers (Event, Scheduled and Tumbling) in ADF.
- Used Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.
- Designed and developed user defined functions, stored procedures, triggers for Cosmos DB
- Analysed the data flow from different sources to target to provide the corresponding design Architecture in Azure environment.
- Created pipelines with GUI in Azure Data factory
- Take initiative and ownership to provide business solutions on time.
- Created High level technical design documents and Application design documents as per the requirements and delivered clear, well-communicated and complete design documents.
- Created DA specs and Mapping Data flow and provided the details to the developer along with HLDs.
- Created Build definition and Release definition for Continuous Integration and Continuous Deployment.
- Created Application Interface Document for the downstream to create a new interface to transfer and receive the files through Azure Data Share.
- Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Data bricks. Created, provisioned different Data bricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters. Integrated Azure Active Directory authentication to every Cosmos DB request sent and demoed feature to Stakeholders
- Improved performance by optimizing computing time to process the streaming data and saved cost to the company by optimizing the cluster run time.
- Created several Databricks Spark Jobs with PySpark to perform several tables to table operations.

- Extensively used SQL Server Import and Export Data tool.
- Created database users, logins and permissions to setup.
- Working with complex SQL, Stored Procedures, Triggers, and packages in large databases from various servers.
- Helping team members to resolve any technical issue, Troubleshooting, Project Risk & Issue identification and management. Addressing resource issue, Monthly one on one, Weekly meeting.
- Creating complicated transformations using Python and Spark SQL in Azure Databricks to implement business rules
- Created applications utilizing Kafka, which tracks consumer lag within Apache Kafka clusters, and used Kafka functions like distribution, partition, and replicated commit log service for messaging systems by maintaining feeds.
- Used Apache Airflow and the Oozie workflow engine to set up a workflow for managing and scheduling Hadoop processes and leveraging the HBase shell and client API to import data into HBase.
- Created various reports using Power BI to display the insights to the business team from the data.
- Developed Storm topologies to read data from Kafka topics, populated staging tables and stored the refined data in partitioned HIVE tables in google cloud services.
- Implemented data access jobs using Hive, Tex, Solr, Base and Storm.

**Environment:** Kafka, PySpark, Spark SQL, YARN, Apache Airflow ELT, Apache Oozie, Power BI, Kubernetes, Helm, Azure, Storm, Teradata, Data Factory, Databricks, HDInsight, SQL, Python, Apache Spark, Airflow, Oozie, Kafka, Python, Spark SQL, ECS fargate

**Client: Comerica Bank, Dallas, TX.** **May 2014 – Aug 2016**
**Role: Data Engineer**

**Responsibilities:**

- Installed, configured, and maintained Apache Hadoop clusters for the development of applications by the specifications.
- Developed T-SQL (DDL, DML, TCL), and SQL Server Programmability Objects such as Stored Procedures, Functions, Triggers, Views, and Sub queries for various business requirements.
- Created Spark programs using Scala and batch processing using functional programming techniques.
- Writing Spark Core Programs to process and clean data before loading it into Hive or HBase to be processed further.
- Created S3 buckets, managing policies for S3 buckets and Glacier for storage and backup on AWS.
- Scheduled Apache Airflow DAGs to export the data to AWS S3 buckets by triggering to invoke an AWS lambda function.
- Collected the data from the edge device databases, exported in CSV format, and stored in AWS S3 buckets.
- Integrated AWS EMR with S3, RedShift and Aurora for ETL.

- Utilization of tools for data transformation such as Data Stage, SSIS, Informatica ELT, or DTS.
- Proficient in using UML for Use Cases, Activity Diagrams, Sequence Diagrams, Data Flow Diagrams, Collaboration Diagrams, and Class.
- Developed robust and scalable data integration pipelines to transfer data from S3 bucket to the Redshift database using Python and AWS Glue.
- In charge of building ETL pipelines with Pig and Hive to extract data from various data sources and import it into the Hadoop Data Lake.
- Used SQL Server Integrations Services (SSIS) to extract, manipulate, and load data from various sources like MOngoDB , MS SQL,etc into the target system.
- Created data mapping, transformation, and cleaning rules for OLTP and OLAP data management. Developed REST APIs using Scala and Play framework to retrieve processed data from Cassandra database.
- Developing UDFs in java for hive and pig and working on reading multiple data formats on HDFS using Scala.
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
- Used Scala collection framework to store and process the complex consumer information.
- Used Scala functional programming concepts to develop business logic. Developed programs in JAVA, Scala-Spark for data reformation after extraction from HDFS for analysis.
- Developed data processing tasks using PySpark such as reading data from external sources, merging the Obtained data, performing data enrichment, and loading into data warehouses.
- Performed the transformations and actions on the imported data from AWS S3 using PySpark.
- Developed Spark scripts by using Scala shell commands as per the requirement.
- Processed the schema oriented and non-schema-oriented data using Scala and Spark.
- Developed Scala scripts, UDFFs using both Data frames/SQL/Data sets and RDD/MapReduce in Spark 1.6 for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
- Provided architecture and design as product is migrated to Scala, Play framework and Sencha UI Implemented applications with Scala along with Akka and Play framework.
- Expert in implementing advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.
- Auction web app - calculated bids for energy auctions utilizing Scala, JPA and Oracle.
- Built Kafka-Spark-Cassandra Scala simulator for MetiStream, a big data consultancy; Kafka-Spark-Cassandra prototypes.
- Experience in AWS cloud infrastructure database migration and converting existing MS SQL to Aurora, MySQL.
- Educate developers on how to commit their work and use CI/CD pipelines that are in place.
- Developed a Restful API using & Scala for tracking open source projects in Github(Bitbucket) and computing the in-process metrics information for those projects.
- Developed analytical components using Scala, Spark, Apache Mesos and Spark Stream.

- Experience in using the Docker container system with the Kubernetes integration Developed a Web Application using Java with the Google Web Toolkit API with Postgresql Redis.
- Creating a dashboard using Flask, Python libraries, and AngularJS to visualize their progress.
- Improve site performance by making better use of caches via Memcached. On Amazon Web Services.
- Used R for prototype on a sample data exploration to identify the best algorithmic approach and then wrote Scala scripts using spark machine learning module.
- Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS.
- Used Tableau for the data visualization during the quick model construction process in Python. These models are then put into practice in SAS, where they are connected to MSSQL databases and have timely update schedules.
- Filter the dataset using PIG UDF/ PIG scripts in HIVE and Storm/Bolt in Apache Storm.
- Utilizing J2EE components such as SOA web services , JSP and Servlets.
- Developed UI layer using Angular JS and HTML according to internal guidelines and standards.
- Worked with DB teams for setting up ASm disks over NAS environments aware, worked with oracle 7000 series from UNIX perspective.
- Developed UNIX shell scripts using Shell scripting .

**Environment:** MapReduce, Spark, Scala, Hive, Pig, Sqoop, Storm,Data Stage, SSIS, Informatica ELT, HBase, Oozie, Impala, Azure, Kafka, JSON, XML Oracle PL/SQL, UML, SQL, HDFS, Unix, Python, PySpark, Power BI, MSSQL, Angular, JSP, CI/CD, Aurora, MongoDB