**Sanket Kumar Patel**

**Sr. Data Engineer**

**P:**

[**Patelsan10091992@gmail.com**](mailto:Patelsan10091992@gmail.com)

**‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬‬**‬‬‬‬‬ ‬‬‬‬‬‬‬‬‬‬‬‬‬‬

**PROFESSIONAL SUMMARY**

* Over 10 years of diversified IT experience in E2E data analytics platforms (ETL - BI-Java) as Bigdata, Hadoop, Java/J2EE Development, Informatica, Data Modeling and System Analysis, In Banking, Finance, Insurance and Telecom domains.
* Worked for 6 years with AWS-Bigdata/Hadoop Ecosystem in the implementation of Data Lake.
* Hands on experience Hadoop framework and its ecosystem like Distributed file system (HDFS), MapReduce, Pig, Hive, Sqoop, Flume, Spark.
* Experience in layers of Hadoop Framework - Storage (HDFS), Analysis (Pig and Hive), Engineering (Jobs and Workflows), extending the functionality by writing custom UDFs.
* Extensive experience in developing Data warehouse applications using Hadoop, Informatica, Oracle, Teradata, MS SQL server on UNIX and Windows platforms and experience in creating complex mappings using various transformations and developing strategies for Extraction, Transformation and Loading (ETL) mechanism by using Informatica.
* Developed AWS Cloud Formation templates and set up Auto scaling for EC2 instances.
* Good knowledge in Apache Hadoop ecosystem components Spark, Cassandra, HDFS, Hive, SQOOP, Airflow.
* Proficient in Hive Query language and experienced in hive performance optimization using Static-Partitioning, Dynamic-Partitioning, Bucketing and Parallel Execution concepts.
* As Data Architect designed and maintained high performance ELT/ETL processes.
* Experience in analyzing data using Hive QL, Pig Latin, and custom MapReduce programs in Java, custom UDF s.
* Good Understanding of Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce concepts.
* Knowledge on Cloud computing infrastructure AWS (amazon web services).
* Created modules for spark streaming in data into Data Lake using Spark.
* Experience in Dimensional Data Modeling Star Schema, Snow-Flake Schema, Fact and Dimensional Tables, concepts like Lambda Architecture, and Batch processing, Oozie.
* Extensively used Informatica client tools Source Analyzer, Warehouse designer, Mapping designer, Mapplet Designer, ETL Transformations, Informatica Repository Manager and Informatica Server Manager, Workflow Manager & Workflow Monitor.
* Experienced in database conversion from Oracle and SQL Server to PostgreSQL and MySQL.
* Good Knowledge on No SQL database like HBase, Cassandra.
* Expertise in using core Java, J2EE, Multithreading, JDBC, Shell Scripting and proficient in using Java API's Collections, Servlets, JSP for application development.
* Worked closely to review pre- and post-processed data to ensure data accuracy and integrity with Dev and QA teams.
* Experience in identifying Bottlenecks in ETL Processes and Performance tuning of the production applications using Database Tuning, Partitioning, Index Usage, Aggregate Tables, Session partitioning, Load strategies, commit intervals and transformation tuning.
* Worked on performance tuning of user queries by analyzing the explain plans, recreating the user driver tables by right Primary Index, scheduled collection of statistics, secondary or various join indexes.
* Experience with scripting languages like PowerShell, Perl, Shell, etc.
* Expert knowledge and experience in fact dimensional modelling (Star schema, Snow flake schema), transactional modelling and SCD (Slowly changing dimension).
* Create clusters in Google Cloud and manage the clusters using Kubernetes(k8s). Using Jenkins to deploy code to Google Cloud, create new namespaces, creating docker images and pushing them to container registry of Google Cloud.
* Experience in Java, J2ee, JDBC, Collections, Servlets, JSP, Struts, Spring, Hibernate, JSON, XML, REST, SOAP Web services, Groovy, MVC, Eclipse, Weblogic, Websphere, and Apache Tomcat severs.
* Working experience with Functional programming languages like Scala, and Java.
* Extensive knowledge of Data Modeling, Data Conversions, Data integration and Data Migration with specialization in Informatica Power Center.
* Expertise in extraction, transformation and loading data from heterogeneous systems like flat files, excel, Oracle, Teradata, MSSQL Server.
* Good work experience with UNIX/Linux commands, scripting and deploying the applications on the servers.
* Strong skills in algorithms, data structures, Object oriented design, Design patterns, documentation and QA/testing.
* Experienced in working as part of fast paced Agile Teams, exposure to testing in scrum teams, Test-Driven development.
* Excellent domain knowledge in Insurance, Telecom and Banking/Finance.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Bigdata Technologies** | AWS EMR, S3, EC2-Fleet, Spark and Hortonworks HDP, Hadoop, Mapreduce, Pig, Hive, Apache Spark, Spark SQL, Informatica Power Canter, Kafka, NoSQL, Elastic Mapreduce (EMR), Hue, YARN, Nifi, Impala, Sqoop, Solr, OOZie. |
| **Databases** | Cloudera Hadoop CDH 15.x, Hortonworks HDP, Oracle, Teradata, DB2, Microsoft SQL Server, MySQL, NOSQL, SQL databases. |
| **Platforms (O/S)** | Red-Hat LINUX, Ubuntu, Windows NT/2000/XP. |
| **Programming languages** | Java, Scala, SQL, UNIX shell script, JDBC, Python, Perl. |
| **Security Management** | Hortonworks Ambari, Cloudera Manager, Apache Knox, XA Secure, Kerberos. |
| **Web-technologies** | DHTML, HTML, XHTML, XML, XSL (XSLT, XPATH), XSD, CSS, JavaScript, SOAP, RESTful, Agile, Design Patterns |
| **Data warehousing** | Informatica PowerCenter/Powermart/Data quality/Bigdata, Pentaho, ETL Development, Amazon Redshift, IDQ. |
| **Database Tools** | JDBC, HADOOP, Hive, No-SQL, SQL Navigator, SQL Developer, TOAD, SQL Plus, SAP Business Objects |
| **Data Modeling** | Rational Rose, Erwin |
| **Code Editors** | Eclipse, IntelliJ |

**[EDUCATIONAL QUALIFICATION]**

**Bachelor of Computer Engineering (INDIA); Gujarat Institute of Technical Studies, 2013**

**[PROFESSIONAL EXPERIENCE**]

### 

### **Cotiviti, Chicago, IL AUG 2021 to present**

### **Sr Data Engineer**

**Responsibilities**:

* Used Hive Context which provides a superset of the functionality provided by SQL Context and Preferred to write queries using the HiveQL parser to read data from Hive tables (fact, syndicate).
* Modeled Hive partitions extensively for data separation and faster data processing and followed Hive best practices for tuning.
* Designed, implemented, and maintained scalable and efficient data engineering architectures within the Databricks platform.
* Develop and execute data quality assessment procedures to evaluate the accuracy, completeness, and consistency of curated datasets.
* Develops Node.js scripts for data processing tasks, such as data extraction, transformation, and loading (ETL) processes.
* Utilizes Node.js to integrate and consume data from various APIs, ensuring seamless connectivity between different data sources.
* Implements streaming data processing using Node.js for real-time data ingestion, transformation, and analysis.
* Developed shell scripts for extracting and transforming healthcare data from various sources, ensuring compatibility and consistency.
* Leveraged Palantir's data integration and analysis platform to develop solutions for managing and processing large datasets.
* Configure and set up AWS Database Migration Service, defining source and target endpoints, replication instances, and migration tasks.
* Utilize AWS DMS to capture and replicate incremental changes (CDC) in source databases to the target, minimizing the impact on system resources.
* Designed and implemented data integration pipelines within Palantir, ensuring seamless ingestion of diverse datasets from various sources.
* Implemented data masking techniques in shell scripts to ensure compliance with healthcare privacy regulations, protecting sensitive patient information.
* Created Docker images for data engineering tools, frameworks, and applications.
* Developed and deployed serverless functions using AWS Lambda for data processing tasks, eliminating the need for managing infrastructure.
* Designed data models and schemas using Palantir's data ontology and entity resolution capabilities to represent complex relationships and hierarchies within datasets.
* Implemented encryption protocols for PII data at rest and in transit, utilizing industry-standard encryption algorithms to safeguard sensitive information.
* Utilized dbt to transform raw data into structured and well-modeled datasets suitable for analytical purposes.
* Developed and maintained dbt projects to organize and structure the data transformation process effectively.
* Orchestrated the deployment of containerized data applications, such as data processing jobs or data pipelines, in Kubernetes clusters.
* Utilized Kubernetes YAML manifests to define and deploy containerized workloads.
* Ensured alignment of the architecture with business requirements and optimized data processing workflows.
* Developed Spark scripts by writing custom RDDs in Scala for data transformations and perform actions on RDDs.
* Caching of RDDs for better performance and performing actions on each RDD.
* Developed robust data ingestion pipelines to efficiently bring in data from diverse sources into the data lake.
* Established and enforced data quality standards and governance policies within the data lake.
* Implemented data profiling and validation processes to ensure the accuracy and integrity of data stored in the data lake.
* Containerized data applications using Docker to encapsulate dependencies and ensure consistency across different environments.
* Created Docker images for various components of data pipelines and processing tasks.
* Implemented data ingestion pipelines to load data into Snowflake from various sources such as databases, data lakes, and external APIs.
* Integrated various data sources with Tableau for comprehensive data analysis and reporting.
* Provisioned and configured EC2 instances to meet the compute and memory requirements of data processing workloads.
* Utilized version control systems for managing code and configuration changes in Big Data projects.
* Optimized EC2 instances for specific data processing tasks by selecting the right combination of instance types, sizes, and storage options.
* Developed and maintained ETL processes to prepare and transform data for optimal use in Tableau.
* Developed and maintained Extract, Transform, Load (ETL) processes using Databricks to process and transform raw data into structured formats.
* Implemented efficient data pipelines for data ingestion, processing, and loading into data lakes or data warehouses.
* Designed and implemented data models for Apache Cassandra, optimizing schema structures for high write throughput and efficient querying.
* Developed and maintained data ingestion pipelines to load and update data into Apache Cassandra from various sources.
* Configured connectivity between different applications and systems using iPaaS connectors, facilitating the integration of data for efficient processing.
* Designed and implemented event-driven architectures using AWS Lambda to respond to changes in data, triggers, or external events.
* Optimized CQL (Cassandra Query Language) queries for performance and efficiency, considering the distributed nature of Cassandra clusters.
* Created Hive Fact tables on top of raw data from different retailer’s which indeed partitioned by Time dimension key, Retailer name, Data supplier name which further processed pulled by analytics service engine.
* Developed highly complex Python and Scala code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
* Implemented real-time data integration solutions using iPaaS, enabling organizations to react promptly to changes in data and ensuring up-to-date information across systems.
* Developed Python scripts to parse XML, Json files and load the data in AWS Snowflake Data warehouse.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, Parquet/Text Files into AWS Redshift.
* Conducted routine maintenance tasks on Kubernetes clusters, including software updates and security patching.
* Developed and maintained ETL processes within Snowflake to transform raw data into a structured and usable format.
* Designed and implemented data integration workflows using SnapLogic to connect diverse data sources and destinations.
* Configured and managed connections between SnapLogic and various data systems, databases, and applications.
* Build a program with Python and Apache beam and execute it in cloud Dataflow to run Data validation between raw source file and Big query tables.
* Building a Scala and spark based configurable framework to connect common Data sources like MYSQL, Oracle, Postgres, SQL Server, Salesforce, Big query and load it in Big query.
* Building data pipeline ETLs for data movement to S3, then to Redshift.
* Designed and implemented ETL pipelines between from various Relational Data Bases to the Data Warehouse using Apache Airflow.
* Implemented data transformation logic within SnapLogic pipelines to ensure data quality and consistency.
* Utilized Apache Spark for distributed data processing within Databricks notebooks, optimizing Spark jobs for performance and scalability.
* Developed custom multi-threaded Java based ingestion jobs as well as Sqoop jobs for ingesting from FTP servers and data warehouses.
* Implemented data quality checks and validations within Databricks to ensure the accuracy and integrity of processed data.
* Writing big query to get data wrangling for with help of data flow in gcp cloud.
* Developed Scala based Spark applications for performing data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
* Programmatically created CICD Pipelines in Jenkins using Groovy scripts, Jenkins file, integrating a variety of Enterprise tools and Testing Frameworks into Jenkins for fully automated pipelines to move code from Dev Workstations to all the way to Prod environment.
* Developed and maintained ETL processes to prepare and transform data for optimal use in Power BI.
* Developed rest API's using python with flask and Django framework and done the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Analyzed SQL scripts and designed the solutions to implement using PySpark.
* Used MongoDB to stored data in JSON format and developed and tested many features of dashboard using Python, Bootstrap, CSS, and JavaScript.
* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Configured and managed data replication strategies in Apache Cassandra to ensure fault tolerance and high availability.
* Developed visually compelling reports and dashboards in Power BI for data exploration and analysis.
* Making a data pipelining with help Data Fabric job, SQOOP, SPARK, Scala and KAFKA. Parallel working in data side oracle and MYSQL server for data designing to source to target.
* Developed various Shell Scripts for scheduling various data cleansing scripts and loading process and maintained the batch processes using Unix Shell Scripts.
* Backing up AWSPostgrestoS3on daily job run onEMRusing Data Frames.
* Developed server-based web traffic using RESTful API's statistical analysis tool using Flask, Pandas.
* Analise various type of raw file like Json, Csv, Xml with Python using Pandas, NumPy etc.
* Created workflows, mappings using Informatica ETL and worked with different transformations such as lookup, source qualifier, update strategy, router, sequence generator, aggregator, rank, stored procedure, filter, joiner, sorter.

**Environments**: HDFS, Hive, AWS, EC2, Airflow, Big Data, Lambda, Sqoop, Spark, Linux, Snowflake, Power BI, Docker, Kafka, Scala, Python, Databricks, Datalake, Stone branch, Cloudera, PySpark, Restful, Oracle, PL/SQL, SQL Server, T-SQL, MYSQL, Unix, Scala, Tableau, Kubernetes

### **Comcast, Dallas, TX FEB 2019 – JULY 2021**

### **Sr Data Engineer/Analyst**

### **Responsibilities:**

* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka and Persists into HDFS.
* Integrated Docker containers with data storage services such as databases, data warehouses, or object storage solutions.
* Managed and maintained Apache Cassandra clusters, including configuration, monitoring, and scaling activities.
* Developed Spark scripts by writing custom RDDs in Scala for data transformations and perform actions on RDDs.
* Used Hive to form an abstraction on top of structured data resides in HDFS and implemented Partitions, Dynamic Partitions, Buckets on HIVE tables.
* Create custom reports and dashboards using business intelligence software’s like Tableau and QuickSight to present data analysis and conclusions.
* Published and maintained workspaces in Power BI Service, allotted the time refresh for the data, and maintained the apps and workbooks.
* Developed custom python program including CI/CD rules for GCP data catalog for metadata management.
* Implemented data quality checks within dbt models to ensure the integrity and consistency of transformed data.
* Used Spark API over Hadoop YARN as execution engine for data analytics using Hive.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.
* Leveraged Palantir's analytical tools and libraries to perform exploratory data analysis, statistical modeling, and machine learning on large-scale datasets.
* Managed configuration files and secrets within Kubernetes for data applications.
* Ensured secure handling of sensitive information, such as database credentials or API keys, through Kubernetes secrets.
* Implemented security measures within iPaaS platforms, including data encryption, access controls, and compliance with security standards to protect sensitive information.
* Utilized monitoring tools provided by iPaaS platforms to track the performance and health of integrations, optimizing configurations for better efficiency.
* Utilized Docker volumes to manage persistent data storage for databases, logs, and other data-intensive applications.
* Integrated diverse data sources with Power BI for comprehensive analysis and reporting.
* Designed, developed data integration programs in a Hadoop environment with NoSQL data store Cassandra for data access and analysis.
* Implemented data security controls and access policies using Palantir's security features to protect sensitive information and ensure compliance with regulatory requirements.
* Implemented data security controls and access policies using Palantir's security features to protect sensitive information and ensure compliance with regulatory requirements.
* Used Tableau to craft data-driven stories, presenting insights in a compelling narrative.
* Configured and managed Snowflake warehouses, ensuring appropriate scaling for concurrent data processing tasks.
* Monitored and managed EC2 costs by selecting cost-effective instance types and sizes.
* Performed schema mapping between source and target data structures, ensuring compatibility during the ingestion process.
* Automated repetitive data engineering tasks using Apache Airflow, reducing manual intervention and improving overall productivity.
* Provided remote Oracle and MySQL database administration and support in Production and Test environments for all Scripps Florida databases. Ensured 100% database availability.
* Managed and maintained Big Data clusters, optimizing configurations for performance and resource utilization.
* Implemented metadata management strategies to catalog and document data assets within the data lake.
* Facilitated easy discovery and understanding of available data resources for data scientists and analysts.
* Used Job management scheduler Apache Oozie to execute the workflow.
* Developed complex mappings using Informatica Power Center Designer to transform and load the data from various source systems like Oracle, Teradata, and Sybase into the final target database.
* Analyzed source data coming different sources like SQL Server tables, XML files and Flat files then transformed according to business rules using Informatica and loaded the data in to target tables.
* Designed and developed a number of complex mappings using various transformations like Source Qualifier, Aggregator, Router, Joiner, Union, Expression, Lookup, Filter, Update Strategy, Stored Procedure, Sequence Generator, etc.
* Involved with Data Analysis Primarily Identifying Data Sets, Source Data, Source Meta Data, Data Definitions and Data Formats.
* Used SSIS to create ETL packages to validate, extract, transform and load data into data warehouse and data marts.
* Collaborated with support teams to ensure the stability and reliability of Big Data systems.
* Implemented Event Handlers and Error Handling in SSIS packages and notified process results to various user communities.
* Configured SQL mail agent for sending automatic emails when SSIS packages fails or succeeds.
* Creating reports using SQL Reporting Services (SSRS) for customized and ad-hoc Queries.
* Created and analyzed business requirements to compose functional and implementable technical data solutions.
* Identified integration impact, data flows and data stewardship.
* Created new data constraints and or leveraged existing constraints for reuse.
* Created data dictionary, Data mapping for ETL and application support, DFD, ERD, mapping documents, metadata, DDL and DML as required.
* Implemented strategies for cost optimization in Snowflake, monitoring resource consumption and optimizing configurations.
* Anticipated JAD sessions as primary modeler in expanding existing DB and developing new ones.
* Involved in creating the Tables in Greenplum and loading the data through Alteryx for Global Audit Tracker.
* Analyzed large and critical datasets using HDFS, HBase, Hive, HQL, PIG, Sqoop and Zookeeper.
* Data Extraction, aggregations, and consolidation of Adobe data within AWS Glue using PySpark.
* Developed Python scripts to automate the ETL process using Apache Airflow and CRON scripts in the UNIX operating system as well.
* Developed data engineering and ETL python scripts for ingestion pipelines which run on AWS infrastructure setup of EMR, S3, Glue and Lambda.
* Changing the existing Data Models using Erwin for Enhancements to the existing Datawarehouse projects.
* Used Talend connectors integrated to Redshift - BI Development for multiple technical projects running in parallel.
* Integrated C++ applications with databases, using APIs such as ODBC or specific database connectors to manage data retrieval and storage.
* Used Presto distributed query engine over hive tables for its high performance and low cost.
* Used custom developed PySpark scripts to pre-process, transform data and map to tables inside the CIF (Non- corporate Information Factory) data warehouse.
* Developed C++ modules for data serialization and deserialization, enabling efficient data storage, transfer, and compatibility across systems.
* Developed shell scripts of Sqoop jobs for loading periodic incremental imports of structured data from various RDMS to S3 and used Kafka to ingest real-time website traffic data to HDFS
* As part of reverse engineering discussed issues/complex code to be resolved and translated them into Informatica logic and prepared ETL design documents.
* Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse.
* Developed mappings in Informatica to load the data from various sources into the Data Warehouse using different transformations like Source Qualifier, Expression, Lookup, aggregate, Update Strategy and Joiner.
* Optimized the performance of the mappings by various tests on sources, targets and transformations.
* Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse.
* Developed mappings in Informatica to load the data from various sources into the Data Warehouse using different transformations like Source Qualifier, Expression, Lookup, aggregate, Update Strategy and Joiner.
* Optimized the performance of the mappings by various tests on sources, targets and transformations.
* Implemented Kerberos for strong authentication to provide data security.
* Implemented LDAP and Active directory for Hadoop clusters.
* Created iterative macro in Alteryx to send Json request and download Json response from webservice and analyze the response data.
* Migrated data from Transactional source systems to Redshift data warehouse using spark and AWS EMR
* Tuning SQL queries using Explain analyzing the data distribution among AMPs and index usage, collect statistics, definition of indexes, revision of correlated sub queries, usage of Hash functions, etc.
* Developed shell scripts for job automation, which will generate the log file for every job.

**Environments***:* Hadoop YARN, Spark, Spark Streaming, Big Data, Spark SQL, MYSQL, Docker, Power BI, Scala, Kafka, Python, Hive, Sqoop, Impala, Tableau, OLTP, Data Lake, Talend, Oozie, Casandra, Control-M, Java, AWSS3, EC2, Airflow, Snowflake, Oracle, Linux, SSIS, SSRS, Data Mapping, Kubernetes, ERD, ETL

### **Bank of America, Charlotte, NC FEB 2017 – JAN 2019**

### **Data Engineer**

### **Responsibilities:**

* Create new UNIX scripts to automate and to handle different file processing, editing and execution sequences with shell scripting by using basic Unix commands and ‘awk’, ‘sed’ editing languages.
* Integrate Collibra with Data Lake using Collibra connect API.
* Utilized Apache Airflow's metadata database to store and manage metadata related to workflow runs, task executions, and dependencies.
* Configured Docker networking to facilitate communication between containerized services.
* Integrated Docker containers into CI/CD pipelines for automated testing, building, and deployment of data applications.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Integrated Lambda functions into microservices architectures to handle specific data processing responsibilities within a larger system.
* Utilized AWS Lambda for real-time data processing scenarios, such as streaming analytics or processing data from IoT devices.
* Maintained comprehensive documentation for Snowflake configurations, data models, and ETL processes.
* Create firewall rules to access Google Data proc from other machines.
* Write Scala program for spark transformation in Dataproc.
* Involved in gathering business requirements, logical modeling, physical database design, data sourcing and data transformation, data loading, SQL, and performance tuning.
* Used SSIS to populate data from various data sources, creating packages for different data loading operations for applications.
* Implemented robust data models within QlikView, ensuring efficient data retrieval, relationships, and optimized performance for complex datasets.
* Integrated QlikView with ETL (Extract, Transform, Load) processes, facilitating seamless data extraction and transformation from various sources.
* Created various types of reports such as complex drill-down reports, drill through reports, parameterized reports, matrix reports, Sub reports, non-parameterized reports and charts using reporting services based on relational and OLAP databases.
* Create Terraform scripts to automate deployment of EC2 Instance, S3, EFS, EBS, IAM Roles, Snapshots and Jenkins Server
* Developed interactive and visually compelling Tableau dashboards for data exploration and analysis.
* Exploring with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, PostgreSQL, Scala, Data Frame, Impala, OpenShift, Talend, pair RDD's.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats.
* Extracted data from various sources like SQL Server, CSV, Microsoft Excel and Text file from Client servers.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Built S3 buckets and managed policies for S3 buckets and used S3 bucket and Glacier for storage and backup on AWS.
* Developed Spark scripts using Python on AWS EMR for Data Aggregation, Validation and Adhoc querying.
* Performed data analytics on DataLake using PySpark on databricks platform
* Involved in creation/review of functional requirement specifications and supporting documents for business systems, experience in database design process and data modeling process.
* Used Kafka producer to ingest the raw data into Kafka topics run the Spark Streaming app to process clickstream events.
* Spark SQL is used as a part of Apache Spark big data framework for structured, Shipment, POS, Consumer, Household, Individual digital impressions, Household TV impressions data processing.
* Created Data Frames from different data sources like Existing RDDs, Structured data files, JSON Datasets, Hive tables, External databases.

Load terabytes of different level raw data into Spark RDD for data Computation to generate the Output response

* Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse.
* Developed mappings in Informatica to load the data from various sources into the Data Warehouse using different transformations like Source Qualifier, Expression, Lookup, aggregate, Update Strategy and Joiner.
* Utilized Airflow to schedule automatically trigger and execute data ingestion pipeline.
* Used Grid Search to evaluate the best hyper-parameters for my model and K-fold cross validation technique to train my model for best results.

**Environments***:* Spark, Redshift, Python, HDFS, EC2, Hive, Sqoop, Big Data, Scala, Databricks, Tableau, Docker, Kafka, Shell scripting, Linux, Jenkins, Eclipse, Git, Oozie, Talend, Agile Methodology, Spark, stone branch, Cloudera, Oracle, PL/SQL, MYSQL, Unix, Json.

### **Nfina Technologies, Alabama MAR 2015 – JAN 2017**

### **Data Engineer**

**Responsibilities:**

* Created and executed Hadoop Ecosystem installation and document configuration scripts on Google Cloud Platform.
* Transformed batch data from several tables containing tens of thousands of records from SQL Server, MySQL, PostgreSQL, and csv file datasets into data frames using PySpark.
* Researched and downloaded jars for Spark-avro programming.
* Developed a PySpark program that writes dataframes to HDFS as avro files.
* Utilized Spark's parallel processing capabilities to ingest data.
* Created and executed HQL scripts that creates external tables in a raw layer database in Hive.
* Developed a Script that copies avro formatted data from HDFS to External tables in raw layer.
* Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, Zookeeper and Sqoop.
* Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
* Evaluated existing infrastructure, systems, and technologies and provided gap analysis, and documented requirements, evaluation, and recommendations of system, upgrades, technologies and created proposed architecture and specifications along with recommendations.
* Installed and Configured Sqoop to import and export the data into Hive from Relational databases.
* Administering large Hadoop environments build and support cluster set up, performance tuning and monitoring in an enterprise environment.
* Close monitoring and analysis of the MapReduce job executions on cluster at task level and optimized Hadoop clusters components to achieve high performance.
* Developed data pipeline using Flume, Sqoop, Pig and Java MapReduce to ingest customer behavioral data into HDFS for analysis.
* Used Python& SAS to extract, transform & load source data from transaction systems, generated reports, insights, and key conclusions.
* Developed story telling dashboards in Tableau Desktop and published them on to Tableau Server which allowed end users to understand the data on the fly with the usage of quick filters for on demand needed information.
* Analyzed and recommended improvements for better data consistency and efficiency
* Designed and Developed data mapping procedures ETL-Data Extraction, Data Analysis and Loading process for integrating data using R programming.
* Integrated HDP clusters with Active Directory and enabled Kerberos for Authentication.

**Environment**:  Hadoop YARN, Spark 1.6, Spark Streaming, Spark SQL, MYSQL, Scala, Pig, Kafka, Python, Hive, Sqoop 1.4.6, Impala, Tableau, Talend, Oozie, Java, AWSS3, Oracle 12c, Linux

### **Ericsson, Atlanta, GA AUG 2013 – FEB 2015**

**Data Analyst**

### **Responsibilities:**

* Analysed business requirements, system requirements, data mapping requirement specifications, and responsible for documenting functional requirements and supplementary requirements in Quality Centre.
* Setting up of environments to be used for testing and the range of functionalities to be tested as per technical specifications.
* Tested Complex ETL Mapping and Sessions based on business user requirements and business rules to load data from source flat files and RDBMS tables to target tables.
* Delivered files in various file formatting systems (ex. Excel file, Tab-delimited text, Coma separated text, Pipe delimited text, etc.)
* Performed ad hoc analyses, as needed, with the ability to comprehend analysis as needed.
* Involved in Teradata SQL Development, Unit testing, and Performance tuning to ensure testing issues are resolved based on using defect reports.
* Tested the database to check field size validation, check constraints, stored procedures and cross-verifying the field size defined within the application with metadata.
* Installed, designed, and developed the SQL Server database.
* Created a logical design of the central relational database using Erwin.
* Configured the DTS packages to run in periodic intervals.
* Extensively worked with DTS to load the data from source systems and run in periodic intervals.
* Worked with data transformations in both normalized and de-normalized data environments.
* Involved in data manipulation using stored procedures and Integration Services.
* Assisted in OLAP and Data Warehouse environment when assigned.
* Created tables, views, triggers, stored procedures, and indexes.
* Designed and implemented database replication strategies for both internal and Disaster Recovery.
* Created ftp connections, and database connections for the sources and targets.
* Maintained security and data integrity of the database.
* Developed several forms & reports using Crystal Reports.

**Environment**, Snowflake, AWS S3, GitHub, Service Now, Map Reduce, EMR, Nebula, Python, Hive, Teradata, SQL Server, Scala, Apache Spark, Sqoop

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*