Professional Summary:

* Over all 9 years of experience in Data Engineer, including profound expertise and experience on statistical data analysis such as transforming business requirements into analytical models, designing algorithms, and strategic solutions that scales across massive volumes of data.
* Experience predominantly on managing the software development of Mortgage, Finance & Lending Management Systems, Retail Banking Solutions & Back Office Automation systems of a leading bank.
* Extensive experience in developing Microservices using Spring Boot, Netflix OSS (Zuul, Eureka, Ribbon, Hystrix) and followed domain driven design.
* Experience in Big Data/Hadoop, Data Analysis, Data Modeling professional with applied information Technology.
* Strong experience in Teradata, Informatica, Python, UNIX shell scripting for processing large volumes of data from varied sources and loading into databases like Teradata, Oracle.
* Excellent Programming skills at a higher level of abstraction using Scala and Java
* Experience in Data Extraction, Transformation and Loading of data from multiple data sources into target databases, using Azure Data bricks, Azure SQL, PostgreSQL, SQL Server, Oracle.
* Confidential Certified Database Administrator/ETL/DW Developer with 10+ years of IT experience in development, analysis, design, testing, presentation and implementation experience using Confidential Redshift, Confidential RDS, SQL Server 2016/2012/2008 R2, Oracle 10g/9i and SQL Server Data tools.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Composer.
* Experienced with cloud platforms like Amazon Web Services, Azure, Data bricks (both on Azure as well as AWS integration of Data bricks).
* Troubleshoot, isolate and correct service-affecting issues on the Network SME in areas including but not limited to: routing protocols, routers, switches, firewall administration, BGP,EIGRP, OSPF VPN
* Experience in Adobe Experience Manager (5.6, 6.0, 6.1, 6.2), Adobe LiveCycle Suite (ES2, ES2.5 and ES4) for cross browser platforms using Adobe LiveCycle Designer (AEM Designer), Acrobat, InDesign development and testing
* Expert in managing data access permissions, ensuring that authorized users have the appropriate level of access while maintaining data security and compliance.
* Extensive experience in Oracle PL/SQL, PostgreSQL pl/pgsql, Microsoft T - SQL.
* Strong experience working with HDFS, Map Reduce, Spark, Hive, Sqoop, Flume, Kafka, Oozie, Pig and HBase.
* Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing Data Mining, Data Acquisition, Data Preparation, Data Manipulation, Feature Engineering, Machine Learning Algorithms, Validation and Visualization and reporting solutions that scales across massive volume of structured and unstructured Data.
* Developing and designing user experiences of Internet/Intranet applications using HTML5, HTML, XHTML XML, CSS, CSS3, JavaScript, jQuery , Backbone JS, Angular JS Meeting W3C standards.
* Hands of experience in GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Pub/suB cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver
* Experience in usage of Hadoop distribution like Cloudera and Hortonworks.
* Excellent Experience in Designing, Developing, Documenting, Testing of ETL jobs and mappings in Server and Parallel jobs using Data Stage to populate tables in Data Warehouse and Data marts.
* Have experience in Apache Spark, Spark Streaming, Spark SQL and NoSQL databases like HBase, Cassandra, and MongoDB.
* Establishes and executes the Data Quality Governance Framework, which includes end - to-end process and data quality framework for assessing decisions that ensure the suitability of data for its intended purpose.
* Expert in designing Server jobs using various types of stages like Sequential file, ODBC, Hashed file, Aggregator, Transformer, Sort, Link Partitioned and Link Collector.
* Proficiency in Big Data Practices and Technologies like HDFS, Map Reduce, Hive, Pig, HBase, Sqoop, Oozie, Flume, Spark, Kafka.
* Expert in designing Parallel jobs using various stages like Join, Merge, Lookup, remove duplicates, Filter, Dataset, Lookup file set, Complex flat file, Modify, Aggregator, XML.
* Extensive experience in Text Analytics, generating data visualizations using R, Python and creating dashboards using tools like Tableau.
* Excellent knowledge of studying the data dependencies using metadata stored in the repository and prepared batches for the existing sessions to facilitate scheduling of multiple sessions.
* Extensive experience in loading and analyzing large datasets with Hadoop framework (Map Reduce, HDFS, PIG, HIVE, Flume, Sqoop, SPARK, Impala, Scala), NoSQL databases like MongoDB, HBase, Cassandra.
* Integrated Kafka with Spark Streaming for real time data processing.
* Experienced on Hadoop Ecosystem and Big Data components including Apache Spark, Scala, Python, HDFS, Map Reduce, KAFKA.
* Experience in using build/deploy tools such as Jenkins, Docker and OpenShift for Continuous Integration & Deployment for Microservices.
* Experience in UNIX shell scripting for processing large volumes of data from varied sources and loading into databases like Teradata and Vertica.
* Implemented Hadoop based data warehouses, integrated Hadoop with Enterprise Data Warehouse systems.
* Hands on experience with big data tools like Hadoop, Spark, Hive, Pig, Impala, PySpark, Spark SQL.
* Good knowledge in Database Creation and maintenance of physical data models with Oracle, Teradata, Natuzzi, DB2, MongoDB, HBase and SQL Server databases.
* Deep understanding of Map Reduce with Hadoop and Spark. Good knowledge of Big Data ecosystem like Hadoop 2.0 (HDFS, Hive, Pig, Impala), Spark (Spark SQL, Spark MLlib, Spark Streaming).
* Experienced in writing complex SQL Quires like Stored Procedures, triggers, joints, and Sub quires.
* Good experience on programming languages Python, Scala.
* Functional Expertise and experience with detailed understanding of the Mortgage and Finance domain helped in understanding and translating the client requirements to innovative scalable solutions to ensure the end product meets the client’s requirements.
* Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales, PivotTables and OLAP reporting.
* Strong experience with Informatica Designer, Workflow Manager, Workflow Monitor, Repository Manager.
* Experience in developing Map Reduce Programs using Apache Hadoop for analyzing the big data as per the requirement. Practical understanding of the Data modeling.
* Provide regular reports on Network SME and VPN performance, capacity, usage, life cycle state, and specific project work impacting network and VPN access.
* Hands on Experience on Query optimization and performance tuning on SQL Server.

**TECHNICAL SKILLS:**

* **Databases:** Oracle, MySQL, SQLite, NO SQL, RDBMS, SQL Server 2014, HBase 1.2, MongoDB 3.2. Teradata, Netezza. Cassandra. Ablation, Data Governance
* **Database Tools:** PL/SQL Developer, Toad, SQL Loader, Erwin.
* **Web Programming:** Html, CSS, Xml, JavaScript.
* **Programming Languages:** R, Python, SQL, Scala, UNIX, C, JAVA, Tableau
* **DWH BI Tools:** Data Stage 9.1, 11.5, Tableau Desktop
* **Data Visualization:** Datalink, Tableau9.4/9.2.
* **Big data Frameworks:** HDFS, Map Reduce, Pig, Hive, Sqoop, Oozie, Zookeeper, Flume and HBase, Amazon EC2, S3 and Red Shift, Spark, Storm, Impala, Kafka.
* **Scheduling Tools:** Autosys, Control-M.
* **Operating Systems:** AIX, LINUX, UNIX.
* **Environment:** AWS, AZURE, Databricks.com
* **Reporting Tools:** SSIS/SSRS/SSAS.

**Client = Anthem Insurance, Houston, TX Mar 2022 to Present**  
**Designation = Lead Data Engineer**  
**Roles & Responsibilities:**

* Used Spark Streaming APIs to perform transformations and actions on the fly for building common learner data model which gets the data from Kafka in Near real time and persist it to Cassandra.
* Consumed XML messages using Kafka and processed the xml file using Spark Streaming to capture UI updates.
* Can work parallels in both GCP and Azure Clouds coherently.
* Designed tables and columns in Redshift for data distribution across data nodes in the cluster keeping columnar database design considerations.
* Provide a level of technical Network SME security expertise to support enterprise network infrastructure.
* Profound knowledge about the architecture of the Teradata database and experience in Teradata Unloading utilities like Fast export.
* Changes/modules. Follow the change control procedures to move the modules into Production Support.
* Created Postgres Tables, Views, Procedures, Functions, Triggers based on the requirements.
* Developed Preprocessing job using Spark Data frames to flatten Json documents to flat file.
* Gathered the business requirements from the Business Partners and Subject Matter Experts
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Managed the implementation, testing, maintenance, and upgrades of all software used by Mortgage Division (NetOxygen, Crystal Report, Powerseller, Infomaker, WorkRequest in Access etc).
* Experience on Cloud Databases and Data warehouses ( SQL Azure and Confidential Redshift/RDS).
* Performed QA/QC for various forms designed in Adobe LiveCycle Designer
* Integrated and accessed data from external sources, including RESTful APIs and web services, to enrich and expand data sets for analysis and application integration.
* Wrote Redshift UDFs and Lambda functions using Python for custom data transformation and ETL.
* Converted a monolithic app to microservices architecture using Spring Boot using 12-factor app methodology. Deployed, Scaled, Configured, wrote manifest file for various Microservices in PCF.
* Designed and implemented a dynamic image rendering web service which is consumed by JSP, jQuery, JavaScript and AJAX.
* Worked on DataStage production support job scheduling process using the Scheduling tools and data stage scheduler.
* Ensured compliance with data access regulations and industry standards, such as GDPR or HIPAA.
* Used Data bricks for encrypting data using server-side encryption.
* Worked on building dashboards in Tableau with ODBC connections from different sources like Big Query/ presto SQL engine.
* Wrote, tested and implemented Teradata Fast load, Multiload and BTEQ scripts, DML and DDL.
* Written multiple Map Reduce programs for data extraction, transformation and aggregation from multiple file formats including XML, JSON, CSV & other compressed file formats.
* Developed automated processes for flattening the upstream data from Cassandra which in JSON format. Used Hive UDFs to flatten the JSON Data.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Developed a Spark Streaming module for consumption of Avro messages from Kafka.
* Worked with different File Formats like TEXTFILE, AVROFILE, ORC, and PARQUET for HIVE querying and processing.
* Worked on Data loading into Hive for Data Ingestion history and Data content summary.
* Involved in developing python scripts, ambition, Informatica and other etl tools for extraction, transformation, loading of data into data warehouse.
* Designing and building multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Confidential Redshift for large scale data handling Millions of records every day.
* Manages and closely monitor the disaster recovery, business continuity, and business resumption plan for the Mortgage Division.
* Maintained a deep understanding of HIPAA regulations and their implications for the organization, ensuring strict compliance with patient data privacy and security requirements.
* Created Impala tables and SFTP scripts and Shell scripts to import data into Hadoop.
* Created Hive tables and involved in data loading and writing Hive UDFs. Developed Hive UDFs for rating aggregation
* Worked on importing and exporting data from snowflake, Oracle and DB2 into HDFS and HIVE using Sqoop for analysis, visualization and to generate reports.
* Used BTEQ and SQL Assistant (Query man) front-end tools to issue SQL commands matching the business requirements to Teradata RDBMS.
* Consult to internal customers regarding application instabilities, Network SME performance problems, and ideas for design improvements.
* Configured and delivered enterprise applications with Adobe Live cycle ES3/4 on a cluster J2EE Environment
* Developed, debugged and optimized PostgreSQL queries for reporting and app development.
* Created data pipeline for different events of ingestion, aggregation and load consumer response data in AWS S3 bucket into Hive external tables in HDFS location to serve as feed for tableau dashboards.
* Used EMR (Elastic Map Reducing) to perform big data operations in AWS.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery.
* Implemented REST Microservices using spring boot. Generated Metrics with method level granularity and Persistence using Spring AOP and Spring Actuator.
* Implementation planning and successful deployment of the developed code in production support environment
* Data Consultant for Confidential working with Google Compute Cloud Data Flow and Big Query to manage and move data within a 200 Petabyte Cloud Data Lake for GDPR Compliance
* Worked on Apache spark writing python applications to convert txt, xls files and parse.
* Developed Python scripts, UDF's using both Data frames/SQL and RDD/Map Reduce in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop.
* Generated documents, LiveCycle processes, maintained document process management using Adobe LiveCycle designer/workbench integrated with J2EE server solution.
* Manages the regulatory requirements and the BAU requirements for the Mortgage division to meet the standards and schedules without failures.
* Implemented Spark using Python/Scala and utilizing Spark Core, Spark Streaming and Spark SQL for faster processing of data instead of Map Reduce in Javascript.
* Involvement in working with Azure cloud stage (HDInsight, Data bricks, Data Lake, Blob, Data Factory, Synapse, SQL DB and SQL DWH).
* Optimizing and tuning the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics
* Involved in running Hadoop Streaming jobs to process Terabytes of data
* Loading data into Snowflake tables from the internal stage using Snow Sql.
* Brought value to the table by redesigning the data entry screens to the Loan Origination System, to simplify the work required by a loan officer working with a customer in applying for a mortgage.

**Environment:** Hadoop, Map Reduce, Hive, HDFS, PIG, Sqoop, Hortonworks, Flume, HBase, Oracle, Snowflake, Teradata, Tableau Unix/Linux, Hadoop, Hive, PIG, SQOOP, Flume, HDFS, Oracle/SQL & DB2, Unix/Linux, JIRA,

**Client = Cotiviti, McKinney TX Jan 2020 to Feb 2022**  
**Designation = Sr. Data Engineer**   
**Roles & Responsibilities:**

* This project was focused on customer clustering. Used the ETL Data Stage Director to schedule and running the jobs, testing and debugging its components & monitoring performance statistics.
* Installed Hadoop, Map Reduce, HDFS, AWS and developed multiple Map Reduce jobs in PIG and Hive for data cleaning and pre-processing.
* Responds to Network SME security incidents. Works along with resource owners during and after security breaches recommends best practices and assesses the scope of damage and risk to vital network resources.
* Migrated on premise database structure to Confidential Redshift data warehouse
* Support and maintain the production support system. Resolve the batch job in Autosys, to bring the system up in
* Conducted HIPAA training sessions for staff, educating them on privacy rules, security measures, and the importance of PHI protection.
* Responsible and accountable for the system study, requirements definition, analysis, design, development and support of the mortgage software used by the bank and the customers.
* I worked with the user community to teach Teradata SQL optimization and intercept poorly performing or highly skewed queries in order to reduce overall impact to the system.
* Implemented Work Load Management (WML) in Redshift to prioritize basic dashboard queries over more complex longer-running adhoc queries.
* Involved in developing the UI pages using HTML, DHTML, CSS, JavaScript, JSON, jQuery, Ajax.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS
* Worked and learned a great deal from AWS Cloud services like EC2, S3, EBS, RDS.
* Created architecture towards a Microservices approach using RESTful Web-Services by using Spring Boot.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, experienced in Maintaining the Hadoop cluster on AWS EMR.
* Worked on AWS Script Conversion Tool (SCT) in converting Oracle schema to PostgreSQL schema.
* Implemented Spark Graph application to analyze guest behavior for data science segments.
* Worked on batch processing of data sources using Apache Spark, Elastic search.
* Implemented the customization of the current mortgage software (Gallagher Financial Millennium) to Web based application (Net Oxygen) by creating scalable design and solutions,
* Worked with data visualization tools including Tableau, Adobe Analytics Omniture, and matplotlib
* Performed information purging and applied changes utilizing Data bricks and Spark information analysis.
* Developed Big Data solutions focused on pattern matching and predictive modeling.
* Collaborated with EDW team in, High Level design documents for extract, transform, validate and load ETL process data dictionaries, Metadata descriptions, file layouts and flow diagrams.
* Provided seamless connectivity between BI tools like Tableau and Qlik to Redshift endpoints.
* Implemented scripts that load Google Big Query data and run queries to export data
* Develop an Estimation model for various product & services bundled offering to optimize and predict the gross margin.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Designed OLTP system environment and maintained documentation of Metadata. Used forward engineering approach for designing and creating databases for OLAP model.
* Designed dynamic client-side JavaScript codes to simulate process for web application, page navigation and form validation.
* Performed duties as a Teradata Database Administrator (DBA) for a 12 node 5350 system. Implemented the Teradata Communications Logical Data Model (cLDM).
* Participates in Information Security Team on call rotation to troubleshoot major Network SME system issues.
* Responsible for Designing Logical and Physical data modelling for various data sources on Confidential Redshift
* Worked on call for production support.
* Implementing or exposing the Microservice architecture with Spring Boot based services interacting through a combination of REST.
* Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* A highly immersive Data Science program involving Data Manipulation & Visualization, Web Scraping, Machine Learning, Python programming, SQL, GIT, Unix Commands, NoSQL, MongoDB, Hadoop.
* Worked on migrating PIG scripts and Map Reduce programs to Spark Data frames API and Spark SQL to improve performance.
* Involved in creating UNIX shell scripts for database connectivity and executing queries in parallel job execution.
* Used the ETL Data Stage Director to schedule and running the jobs, testing and debugging its components & monitoring performance statistics.
* Led incident response efforts in the event of a data breach or security incident, ensuring proper reporting, investigation, and mitigation in compliance with HIPAA requirements.
* Worked closely with the ETL Developers in designing and planning the ETL requirements for reporting, as well as with business and IT management in the dissemination of project progress updates, risks, and issues.
* Performed scoring and financial forecasting for collection priorities using Python, and SAS.
* Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
* Data Extraction, aggregations and consolidation of Adobe data within AWS Glue using PySpark
* Worked on data migration (full load and change data capture (CDC)) to AWS Amazon Aurora/ RDS with PostgreSQL compatibility using Data Migration Services (DMS).
* Worked in AWS environment for development and deployment of custom Hadoop applications.

**Client = Blue yonder,** **Coppell TX Jan 2018 to Dec 2019**  
**Designation = Data Engineer**  
**Roles & Responsibilities:**

* Responsible for analyzing large data sets to develop multiple custom models and algorithms to drive innovative business solutions.
* Lead the team for the support & development of mortgage software for Confidential.
* Involved in designing data warehouses and data lakes on regular (Oracle, SQL Server) high performance on big data (Hadoop - Hive and HBase) databases. Data modeling, Design, implement, and deploy high-performance, custom applications at scale on Hadoop /Spark.
* Implemented Micro Services based Cloud Architecture on AWS Platform.
* Build a program with Python sdk with Apache beam framework and execute it in Cloud Dataflow to stream pub sub messages into big query tables.
* Maintain a thorough understanding of (Network SME services, (i.e., IPAM, DNS/DHCP, NTP, SNMP), Network Security, IP Routing, VPN,
* Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift
* Generated ad-hoc SQL queries using joins, database connections and transformation rules to fetch data from legacy DB2 and SQL Server database systems.
* Solving the production issues within SLA timelines.
* Implement complex ETL procedures for TeraData Data Warehouse using Unix (Korn Shell scripts, awk, and sed), PL/SQL, bteq, fastload and other utilities.
* Got involved in migrating on prem Hadoop system to using GCP (Google Cloud Platform).
* Translated business requirements into working logical and physical data models for OLTP &OLAP systems.
* Creation of BTEQ, Fast export, Multi Load, Pump, Fast load scripts for extracting data from various production systems.
* Reviewed Stored Procedures for reports and wrote test queries against the source system (SQL Server-SSRS) to match the results with the actual report against the DataMart (Oracle)
* Overloading of functions is performed in Postgres DB based on different number of input parameters.
* Perform Data profiling, preliminary data analysis and handle anomalies such as missing, duplicates, outliers, and imputed irrelevant data. Remove outliers using Proximity Distance and Density based techniques.
* Involved in Analysis, Design and Implementation/translation of Business User requirements.
* Used supervised, unsupervised and regression techniques in building models.
* Performed Market Basket Analysis to identify the groups of assets moving together and recommended the client their risks.
* Participated in Mortgage Insurance (MI) Loan origination, Pricing and selling, Underwriting system interfaces.
* Used JavaScript and XML to update a portion of a web page thus reducing bandwidth usage and load
* Developed ETL (Extraction, Transformation and Loading) procedures and Data Conversion Scripts using Pre-Stage, Stage, Pre-Target and Target tables.
* Maintained thorough documentation of HIPAA compliance efforts, risk assessments, policies, and incident response plans.
* Creating the data pipelines using state of the art Big Data frameworks/tools
* Worked on storing the data frame into hive as table using Python (PySpark).
* Deployed Spring Boot based micro services Docker container using Amazon EC2 and using AWS admin console.
* Hands on experience in developing apache SPARK applications using Spark tools like RDD transformations, Spark core, Spark MLlib, Spark Streaming and Spark SQL.
* Implemented Partitioning, Dynamic Partitions, Buckets in Hive.
* Build a program using Python and Apache beam to execute it in cloud Dataflow and to run Data validation jobs between raw source file and big query tables.
* Used JSON schema to define table and column mapping from S3 data to Redshift
* Utilizing Spark get data from HDFS, process it and store it back into HDFS
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3 buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Deployed Microservices/Cloud Native App's in Pivotal Cloud Foundry (Paas).
* Created Airflow Scheduling scripts in Python to automate the process of Sqooping wide range of data sets.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS
* Developed Spark SQL scripts using Python for faster data processing
* Using Sqoop to extract the data from warehouse, SQL server and load into Hive
* Used Spark framework to transform the data for final consumption of analytical applications
* Involved in scheduling Oozie workflow engine to run multiple Hive jobs, developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Spark.
* Responsible for ensuring that Confidential &T Wi-Fi Network SME delivers extremely high availability while scaling with the needs of the business.
* Performed Exploratory Data Analysis using R. Also involved in generating various graphs and charts for analyzing the data using Python Libraries.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Dynamic implementation of SQL server work on website using SQL developer tool and Experience with continuous integration and automation using Jenkins and Implemented Service Oriented Architecture (SOA) using JMS for sending and receiving messages while creating web services.

**Client = V-Soft , Louisville KY Nov 2015 to Nov 2017  
Designation = Data Engineer**  
**Roles & Responsibilities:**

* Involved in Analysis, Design and Implementation/translation of Business User requirements.
* Worked on collection of large sets using Python scripting. Spark SQL
* Worked on large sets of Structured and Unstructured data.
* Worked on creating DL algorithms using LSTM and RNN.
* Actively involved in designing and developing data ingestion, aggregation, and integration in Hadoop environment
* Developed SQOOP scripts to migrate data from Oracle to Big Data Environment.
* Converted all Hadoop jobs to run in EMR by configuring the cluster according to the data size
* Implemented Spring security for SQL injunction and user access privileges, Used various Java, J2EE design patterns like DAO, DTO, Singleton etc.
* Solves problems causing Network SME outages including identifying issues with hardware, operating system, application and capacity.
* Performed data analysis and data profiling using complex SQL queries on various sources systems including Oracle 10g/11g and SQL Server 2012.
* Identified inconsistencies in data collected from different source.
* Participated in requirement gathering and worked closely with the architect in designing and modeling.
* Designed object model, data model, tables, constraints, necessary stored procedures, functions, triggers, and packages for Oracle Database.
* Wrote Spark applications for Data validation, cleansing, transformations and custom aggregations.
* Imported data from various sources into Spark RDD for processing.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Worked on installing cluster, commissioning & decommissioning of Data node, Name node high availability, capacity planning, and slots configuration.
* Developed Spark applications for the entire batch processing by using Scala.
* Automatically scale-up the EMR instances based on the data.
* Stored the time-series transformed data from the Spark engine built on top of a Hive platform to Amazon S3 and Redshift.
* Facilitated deployment of multi-clustered environment using AWS EC2 and EMR apart from deploying Dockers for cross-functional deployment.
* Visualized the results using Tableau dashboards and the Python Seaborn libraries were used for Data interpretation in deployment.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Sql Activity.
* Involved in running Hadoop Streaming jobs to process Terabytes of data
* Loading data into Snowflake tables from the internal stage using Snow Sql.
* Build the logical and physical data model for Snowflake as per the changes required.
* Created Business-Crucial stored procedures, views, and functions to support efficient data storage and manipulation.

**Client = Fifth third bank,** **Sanjose CA Sep 2014 to Oct 15  
Designation = Data Scientist**  
**Roles & Responsibilities:**

* Worked on different dataflow and control flow task, for loop container, sequence container, script task, executes SQL task and Package configuration.
* Performance tuning the tables in Redshift.
* Created new procedures to handle complex logic for business and modified already existing stored procedures, functions, views and tables for new enhancements of the project and to resolve the existing defects.
* Loading data from various sources like OLEDB, flat files to SQL Server 2012 database Using SSIS Packages and created data mappings to load the data from source to destination.
* Created batch jobs and configuration files to create automated process using SSIS.
* Created SSIS packages to pull data from SQL Server and exported to Excel Spreadsheets and vice versa.
* Built SSIS packages, to fetch file from remote location like FTP and SFTP, decrypt it, transform it, mart it to data warehouse and provide proper error handling and alerting
* Data validation and cleansing of staged input records was performed before loading into Data Warehouse