**Senior Analytics Engineer**

**Midhun Mallela**

**Email:****huntermax862@gmail.com**

**Ph: +1 8155569034**

**PROFESSIONAL SUMMARY**

* Overall 8+ years of professional experience with **Analytics Engineer**, Azure, and SQL consultant experience in ingestion, data modeling, querying, processing, storage, analysis, Data Integration, and implementing enterprise-level systems. Responsibility, Timely, Teamwork, leadership, and documentation are inherited for the success of any given project.
* Experience in design and deployment of Enterprise Application Development, Web Applications, Client-Server Technologies, and Web Programming using Java and Big data technologies.
* Expertise in Hadoop architecture and ecosystem such as HDFS, MapReduce, Pig, Hive, Sqoop Flume, and Oozie.
* Complete Understanding of Hadoop daemons such as Job Tracker, Task Tracker, Name Node, Data Node, and MRV1 and YARN architecture.
* Experience in installation, configuration, management, supporting, and monitoring Hadoop clusters using various distributions such as Apache Hadoop, Cloudera Hortonworks, and various cloud services like AWS, GCP.
* Good knowledge in Database Creation and maintenance of physical data models with Oracle, Teradata, Netezza, DB2, MongoDB, HBase and SQL Server databases.
* Hands-on experience with AWS services, such as EMR, Kinesis stream, kinesis firehose, IAM, S3, AWS sage maker, EC2, route S3, RDS, Elastic Load balancer ELB, DynamoDB, Glue, SNS, SQS, Cloud formation and have pretty good experience on Redshift spectrum and AWS Athena query services for reading the data from S3.
* Have Experience in IT data analytics projects, Firsthand experience in migrating on-premises ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud DataProc, Google Cloud Storage, Composer.
* Firsthand experience in MS SQL Server with Business Intelligence in SQL Server Integration Services (SSIS), SQL Server Analysis Services (SSAS), SQL Server Reporting Services (SSRS), Azure Cloud Technologies including Azure Database, Azure SQL, Azure Datawarehouse, Azure Data Factory (ADF), Azure Data Lake (ADL), Azure Databricks (ADB).
* Integrated Kafka with Spark Streaming for real time data processing.
* Strong experience with spark real time streaming data using Kafka and Spring boot API.
* 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.
* Understanding of structured data sets, data pipelines, ETL tools, data reduction, transformation and aggregation technique, Knowledge of tools such as DBT, DataStage.
* Experience in Installation and Configuring Hadoop Stack elements MapReduce, HDFS, Hive, Pig, Sqoop, Flume, Oozie, and Zookeeper.
* Expertise in writing custom Kafka consumer code and modifying existing producer code in Python to push data to Spark-streaming jobs.
* Ample knowledge on Apache Kafka, Apache Storm to build data platforms, pipelines, and storage systems; and search technologies such as Elastic search.
* Experience in using Elastic search to store, search, and analyze big volumes of data quickly and in near real-time.
* Proficient with complex workflow orchestration tools namely Oozie, Airflow, Data pipelines and Azure Data Factory, CloudFormation & Terraforms.
* Implemented Data warehouse solution consisting of ETLS, On-premises to Cloud Migration and good expertise building and deploying batch and streaming data pipelines on cloud environments.
* Worked on Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and familiar with building custom Airflow operators and orchestration of workflows with dependencies involving multi-clouds.
* Leveraged Spark as ETL tool for building data pipelines on various cloud platforms like AWS EMRS, Azure HD Insights and MapR CLDB architectures.
* Career Interest and future aspirations include but are not limited to: ML, AI, RPA & Automation everywhere motives.
* Spark for ETL follower, Databricks Enthusiast, Cloud Adoption & Data Engineering enthusiast in Open-source community.
* Proven expertise in deploying major software solutions for various high-end clients meeting business requirements such as Big Data Processing, Ingestion, Analytics and Cloud Migration from On-prem to Cloud.
* Proficient with Azure Data Lake Services (ADLS), Databricks &python Notebooks formats, Databricks Delta lakes & Amazon Web Services (AWS).
* Orchestration experience using Azure Data Factory, Airflow 1.8, and Airflow 1.10 on multiple cloud platforms and able to understand the process of leveraging the Airflow Operators.
* Developed and Deployed various Lambda functions in AWS with in-built AWS Lambda Libraries and deployed Lambda Functions in Scala with custom Libraries.
* Experience new features implemented by Azure to reproduce and troubleshoot Azure end-user issues and provide solutions to mitigate the issue.
* Knowledge in automated deployments leveraging Azure Resource Manager Templates, DevOps, and Git repository for Automation and usage of Continuous Integration (CI/CD).
* Experienced in data processing and analysis using Spark, HiveQL, and SQL.
* Experience in Writing User-Defined Functions (UDFs) in Hive and Spark.
* Worked on Apache Sqoop to perform importing and exporting data from HDFS to RDBMS/NoSQL DBs and vice-versa.
* Working experience on NoSQL databases like HBase, Azure, MongoDB, and Cassandra with functionality and implementation.
* Worked extensively over semi-structured data (fixed length & delimited files) for data sanitation, report generation, and standardization.
* Experience working with AWS Cloud services and AWS SDKs to work with services like AWS API Gateway, Lambda, S3, IAM, and EC2.
* Experience working with GCP Cloud services like GCP cloud dataflow, GCP Big Query, GCP Cloud Storage, GCP PUB/SUB
* Exceptional understanding of Zookeeper for monitoring and managing Hadoop jobs.
* Experience with NumPy, Matplotlib, Pandas, Seaborn, and Matplot Python libraries. Worked on large datasets by using PySpark, NumPy, and pandas.
* Utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Experience in HANA security including User Management, Roles, and Analytic Privileges.
* Experience in HANA SQL Script Stored Procedures, Table functions, and dynamic privileges by SQL query in information models.
* Experience in using Kafka and Kafka brokers to initiate spark context and processing live streaming.
* Experience across both relational databases and non-relational databases Oracle, PL/SQL, SQL Server, MySQL, and DB2.
* Used Project Management services like JIRA for handling service requests and tracking issues. Valuable experience with Software methodologies like Agile and Waterfall. Experienced working with Zookeeper to provide coordination services to the cluster.
* Skilled in Tableau 9 for data visualization, Reporting and Analysis. Good collaborator with ability to solve problems, organize and prioritize multiple tasks.
* Holds a strong ability to manage multiple priorities and workloads and can understand and adapt to modern technologies and environments faster.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, Yarn, MapReduce, Spark, Kafka, Kafka Connect, Hive, Airflow, Stream Sets, Sqoop, HBase, Flume, Pig, Ambari, Oozie, Zookeeper, Nifi, Sentry |
| **Hadoop Distributions** | Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP |
| **Cloud Environment** | Amazon Web Services (AWS), Microsoft Azure |
| **Databases** | MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2 |
| **NoSQL Database** | DynamoDB, HBase |
| **AWS** | EC2, EMR, S3, Redshift, EMR, Lambda, Kinesis Glue, Data Pipeline |
| **Microsoft Azure** | Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Azure Active Directory |
| **Operating systems** | Linux, Unix, Windows 10, Windows 8, Windows 7, Windows Server 2008/2003, Mac OS |
| **Software’s/Tools** | Microsoft Excel, Stat graphics, Eclipse, Shell Scripting, ArcGIS, Linux, Jupyter Notebook, PyCharm, Vi / Vim, Sublime Text, Visual Studio, Postman, Ansible |
| **Reporting Tools/ETL Tools** | Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, Power BI, Arcadia, Data stage, Pentaho |
| **Programming Languages** | Python (Pandas, SciPy, NumPy, Scikit-Learn, Stats Models, Matplotlib, Seaborn, Keras, TensorFlow, PyTorch), PySpark, T-SQL/SQL, PL/SQL, HiveQL, Scala, UNIX Shell Scripting |
| **Version Control** | Git, SVN, Bitbucket |
| **Development Tools** | Eclipse, NetBeans, IntelliJ, Hue, Microsoft Office |

**WORK EXPERIENCE**

**SR. Analytic Engineer**

**Client Name: Bank of America, Chicago, Illinois**

**Duration:2023 June to present**

**Responsibilities:**

* Developed deep understanding of the data sources, implemented data standards,maintained data quality, and mastered data management.
* Used T-SQL in constructing User Functions, Views, Indexes, User Profiles, Relational Database Models, Data Dictionaries, and Data Integrity.
* Expert in building Databricks notebooks, extracting the data from various source systems like DB2, Teradata and perform data cleansing, data wrangling, data ETL processing and loading to AZURE SQL DB.
* Expert in building Ephemeral Notebooks in Databricks like wrapper, driver, and confit for processing the data, back feeding the data to DB2 using multiprocessing thread pool.
* Expert in developing JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that processes the data.
* Extensively used SQL Server Import and Export Data tool.
* Created database users, logins, and permissions to setup.
* Loaded data from various sources like flat files, to SQL Server database Using SSIS Package.
* Implemented Side - by- Side Migration of MS SQL SERVER 2016.
* Involved in daily production server check list, SQL Backups, Disk Space, Job Failures, System Checks, checking performance statistics for all servers using monitoring tool and research and resolve any issues, checking connectivity.
* Established a formal EDM, MDM (Master Data Management) program that creates effective engagement between Business operation, EDM, delivery team and IT.
* Designed and developed SSIS (ETL) packages to validate, extract, transform and load data from OLTP system to the Data warehouse.
* Designed and implemented Tables, Functions, Stored Procedures and Triggers in SQL Server 2016 and wrote the SQL.
* Expert in using Databricks with Azure Data Factory (ADF) to compute large volumes of data.
* Performed ETL operations in Azure Databricks by connecting to different relational database source systems using job connectors.
* Developed Python scripts to do file validations in Data bricks and automated the process using ADF.
* Analyzed the SQL scripts and designed it by using Pyspark SQL for faster performance.
* Worked on reading and writing multiple data formats like JSON, Parquet, and delta from various sources using Pyspark.
* Developed an automated process in Azure cloud, which can ingest data daily from web service and load into Azure SQL DB.
* Experience in building efficient pipeline for moving data between GCP and Azure using Azure Data Factory
* Experience in designing and implementing data processing solutions using Azure Databricks.
* Proficient in working with Medallion architecture to build scalable and efficient data pipelines.
* Developed multi cloud strategies in better using GCP (for its PASS) and Azure (for its SAAS).
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Create Data Flows of existing SSIS Packages in Azure Data Factory.
* Designing and maintaining ADF pipelines with activities Copy, Lookup, For Each, Get Metadata, Execute Pipeline, Stored Procedure, if condition, Web, Wait, Delete etc.
* Captured SCD2 and updated, inserted, or deleted based on Business requirement using Databricks.\
* Worked on Ingestion of data from source (On prem SQL) to target (ADLS) using Azure Synapse Analytics (DW) & Azure SQL DB.
* Develop up the Framework for creation of new snapshots and deletion of old snapshots in Azure Blob Storage and set up the life cycle policies to back the data from delta lakes.
* Expert in building the Azure Notebooks functions by using Python, Scala, and Spark.
* Integrated both framework and CloudFormation to automate Azure environment creation along with ability to deploy on Azure, using build scripts (Azure CLI) and automated solutions using terraform.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Understanding the Business requirements and developing common solutions that meet the business requirement.
* Conducted quarterly Data owner meetings to communicate upcoming Data Governance initiatives, processes, policies, and best practices.
* 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 Databricks 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 company by optimizing the cluster run time. Perform ongoing monitoring, automation and refinement of data engineering solutions prepare complex SQL views, stored procedures in Azure SQL Datawarehouse and Hyperscale.
* Developed reusable Terraform modules to standardize infrastructure deployments.
* Implemented Terraform pipelines for automated infrastructure provisioning and management.
* Designed and developed a new solution to process the NRT data by using Azure stream analytics, Azure Event Hub, and Service Bus Queue. Created Linked service to land the data from SFTP location to Azure Data Lake.
* Automated the build, testing, and deployment of data engineering artifacts using Azure CI/CD tools.
* Integrated Azure Databricks and Delta Lake pipelines into CI/CD workflows for seamless deployment and monitoring.
* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks etc.
* Created several Databricks Spark jobs with PySpark to perform several tables to table operations.
* Working with complex SQL, Stored Procedures, Triggers, and packages in large databases from various servers.
* Writing a Data Bricks code and ADF pipeline fully parameterized for efficient code management.
* Strong skills in visualization tools Power BI, Confidential Excel - formulas, Pivot Tables, Charts and DAX Commands.
* Created few Power BI dashboard reports, Heat map charts and supported numerous dashboards, pie charts and heat map charts that were built on Teradata database.
* Implement Copy activity, Custom Azure Data Factory Pipeline Activities.
* Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL.
Collaborate with application architects on infrastructure as a service (IaaS) application to Platform as a Service (PaaS).
* Helping team members to resolve any technical issue, Troubleshooting, Project Risk & Issue identification, and management Addressing resource issue, Monthly one on one, Weekly meeting.
* Optimized Delta Lake performance by leveraging features such as compaction, caching, and indexing.
* Implemented Delta Lake CDC (Change Data Capture) to capture and process real-time data changes.

**Environment:** Azure SQL Database, Azure Data Lake, Azure Data Factory (ADF), Azure SQL Data Warehouse, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, Azure Data Factory (ADF), Azure Database Migration Service (DMS), GIT, PySpark, Python, JSON, ETL Tools, SQL Azure.

**SR. Analytics Engineer**

**Client Name: ABC Supply, Chicago, Illinois**

**Duration:2022 August to 2023 May**

**Responsibilities:**

* Develop deep understanding of the data sources, implement data standards, maintain data quality, and master data management.
* Expert in building Databricks notebooks, extracting the data from various source systems like DB2, Teradata and perform data cleansing, data wrangling, data ETL processing and loading to AZURE SQL DB.
* Expert in building Ephemeral Notebooks in Databricks like wrapper, driver, and confit for processing the data, back feeding the data to DB2 using multiprocessing thread pool.
* Expert in developing JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that processes the data.
* Expert in using Databricks with Azure Data Factory (ADF) to compute large volumes of data.
* Performed ETL operations in Azure Databricks by connecting to different relational database source systems using job connectors.
* Developed Python scripts to do file validations in Data bricks and automated the process using ADF.
* Integrated data pipelines with external systems and APIs for data extraction and loading.
* Implemented scalable and distributed data processing using tools like Apache Spark and PySpark.
* Analyzed the SQL scripts and designed it by using Pyspark SQL for faster performance.
* Worked on reading and writing multiple data formats like JSON, Parquet, and delta from various sources using Pyspark.
* Experience in building data pipelines and ETL workflows using Python and related libraries (e.g., Pandas, NumPy, SQL Alchemy).
* Developed an automated process in Azure cloud, which can ingest data daily from web service and load into Azure SQL DB.
* Designing and maintaining ADF pipelines with activities Copy, Lookup, For Each, Get Metadata, Execute Pipeline, Stored Procedure, if condition, Web, Wait, Delete etc.
* Captured SCD2 and updated, inserted, or deleted based on Business requirement using Databricks.
* Integrated Azure Databricks and Delta Lake pipelines into CI/CD workflows for seamless deployment and monitoring.
* Implemented version control and branching strategies for data engineering code repositories.
* Automated data pipeline monitoring and alerting using Azure CI/CD and monitoring tools.
* Worked on Ingestion of data from source (On prem SQL) to target (ADLS) using Azure Synapse Analytics (DW) & Azure SQL DB.
* Developed the Framework for creation of new snapshots and deletion of old snapshots in Azure Blob Storage and set up the life cycle policies to back the data from delta lakes.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from various sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Designed and optimized Databricks clusters for improved performance and cost efficiency.
* Developed and implemented automation scripts for managing Databricks workspaces and clusters.
* Connected Azure SQL to Power BI and Built Power BI reports using SQL Azure as data source.
* Developed complex SQL queries using stored procedures, common table expressions (CTEs), temporary tables to support Power BI and SSRS reports.
* Embedded Power BI reports on SharePoint portal page and managed access of reports and data for individual users using Roles.
* Implemented schema evolution and versioning strategies for Delta Lake tables.
* Optimized Delta Lake performance by leveraging features such as compaction, caching, and indexing.
* Successfully redesigned and migrated complex applications from Access to SharePoint and SQL Server 2008 by evaluating database structures and data movements.
* Developed and maintained multiple Power BI dashboards/reports and content packs.
* Preparing necessary pipelines to trigger containers with Azure Data Factory
* Proficient in Infrastructure as Code (IaC) concepts and implemented infrastructure provisioning using Terraform.
* Designed and managed Azure resources (e.g., virtual machines, storage accounts, networks) using Terraform scripts.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that processes the data using the SQL Activity.
* Involving in monitoring, Scheduling and Managing the workflows at Apache Airflow and
* Managing the storage in Azure Data Lake storage (ADLS) by delivering the processed data using Highly optimized storage techniques.
* Expert in building the Azure Notebooks functions by using Python, Scala, and Spark.
* Built and configured a virtual data Centre in the Azure cloud to support Enterprise Data Warehouse hosting including Virtual Private Cloud (VPC), Public and Private Subnets, Security Groups, Route Tables.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Writing Data Bricks code and ADF pipeline fully parameterized for efficient code management.
* Developed spark applications in python (PySpark) on the distributed environment to load enormous numbers of CSV files with different schema into Hive ORC tables.
* Collaborated with cross-functional teams to gather requirements and design data engineering solutions.
* Developed and maintained data engineering documentation, including design specifications and technical documentation.
* Find and resolve complex build and deployment issues while on-call. 24\*7 support and strong knowledge of troubleshooting and debugging application team issues.

**Environment:** Azure SQL Database, Azure Data Lake, Azure Data Factory (ADF), Azure SQL Data Warehouse, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, Azure Data Factory (ADF), Azure Database Migration Service (DMS), GIT, PySpark, Python, JSON, ETL Tools, SQL Azure.

**SR. Data Engineer**

**Client Name: Experian, CA**

**Duration: 2019 Aug to 2022June**

**Responsibilities:**

* Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and Step Functions.
* Created yaml files for each data source and including glue table stack creation.
* Worked on a python script to extract data from Netezza databases and transfer it to AWS S3
* Developed Lambda functions and assigned IAM roles to run python scripts along with various triggers (SQS, Event Bridge, SNS)
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.
* Design & develop the process to get data from multiple environments and create a Data Lake in AWS Environment
* Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and Step Functions.
* Created a Lambda Deployment function and configured it to receive events from S3 buckets.
* Migrated data from on-premises to AWS storage buckets using AWS Data Sync.
* Developed Mappings using Transformations like Expression, Filter, Joiner, and Lookups for better data messaging and to migrate clean and consistent data.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Developed a python script to transfer data from on-premises to AWS S3.
* Designed and deployed ML model to prod which involves pre and post Data Engineering work using serverless components like lambda, Sage maker (Processing job), Step Function, Cloud watch (Event Bridge).
* Maintain proficiency in AWS coding standards and Reporting tools, program development, testing and serve as an expert in area of responsibility.
* Collected data using Spark Streaming from AWS S3bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
* Responsible for building scalable distributed data solutions using the EMR cluster environment with Amazon EMR.
* Migrate data from on-premises to AWS storage buckets.
* Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Installed/Configured/Maintained Apache Hadoop clusters for application development based on the requirements.
* Installed and configured Hadoop Map Reduce, HDFS, developed multiple Map Reduce jobs in java and Scala for data cleaning and preprocessing.
* Developed Spark scripts by using Scala Shell commands as per the requirement.
* Developed Spark code using python for PySpark and Spark-SQL for faster testing and processing of data.
* Created data pipelines for different events to load the data from Dynamo DB to AWS S3 bucket and then into HDFS location.
* Performed Data Cleaning, features scaling, features engineering using pandas and numpy packages in python.
* Created interactive and highly informative Power BI reports.
* Involved in importing real time data to Hadoop using Kafka and implemented the Oozie job for daily imports.
* Developed a Spark job which indexes data into Elastic Search from external Hive tables which are in HDFS.
* Developed Spark programs with Scala and applied principles of functional programming to do batch processing.
* Involved in the development of real time streaming applications using PySpark, Apache Flink, Kafka, Hive on distributed Hadoop Cluster.
* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto- scaling in AWS Cloud Formation.
* Developed application to clean semi-structured data like JSON/XML into structured files before ingesting them into HDFS.
* Built real time pipeline for streaming data using Kafka and Spark Streaming.
* Writing HiveQL as per the requirements and Processing data in Spark engine and store in Hive tables.
* Responsible for importing data from Postgres to HDFS, HIVE using SQOOP tool.
* Developed stored procedures/views in Snowflake and use in Talend for loading Dimensions and Facts.
* Writing the Spark Core Programs for processing and cleansing data thereafter load that data into Hive or HBase for further processing.
* Developed framework for automated data ingestion from various sources like relational databases, delimited files, JSON files, XML files into HDFS and build Hive/Impala tables on top of them.
* Developed real-time data ingestion application using Flume and Kafka.
* Hive tables were created on HDFS to store the data processed by Apache Spark on the Cloudera Hadoop Cluster in Parquet format.
* Developed a tool to load S3 JSON file into Hive table in parquet format in Scala and Apache Spark.
* Written a tool that scrubs numerous files in Amazon S3, getting rid of unwanted characters and other activities using Scala.

**Environment:** Cloudera, Hive, Impala, Spark, Apache Kafka, Flume, Scala, AWS, EC2, S3, Dynamo DB, Auto Scaling, Lambda, Nifi, Power BI, Snowflake, Java, Shell-scripting, SQL, Sqoop, Oozie, Java, PL/SQL, Oracle 12c, SQL Server, HBase,

**Data Analyst**

**Client Name: CITI Bank, NCR**

**Duration: 2018 July to 2019 Aug**

**Responsibilities:**

* Actively Participated in all phases of the Software Development Life Cycle (SDLC) from implementation to deployment.
* Responsible for building scalable distributed data solutions using Hadoop.
* Responsible for Cluster maintenance, adding and removing cluster nodes, Cluster Monitoring and Troubleshooting, Managing, and reviewing data backups & log files.
* Responsible for managing the test data coming from various sources.
* Analyzed data using Hadoop components Hive and Pig.
* Load and transform large sets of structured, semi structured, and unstructured data using Hadoop/Big Data concepts.
* Involved in importing and exporting the data from RDBMS to HDFS and vice versa using Sqoop.
* Developed ETL processes (Data Stage Open Studio) to load data from multiple data sources to HDFS using FLUME and SQOOP, and performed structural modifications using Map Reduce, HIVE.
* Developing Spark scripts, UDFS using both Spark DSL and Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.
* Strong understanding of Partitioning, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Interacted with business partners, Business Analysts, and product owners to understand requirements and build scalable distributed data solutions using the Hadoop ecosystem.
* Developed Spark Streaming programs to process near real time data from Kafka, and process data with both stateless and stateful transformations.
* Experience in report writing using SQL Server Reporting Services (SSRS) and creating distinct types of reports like drill down, Parameterized, Cascading, Conditional, Table, Matrix, Chart and Sub Reports.
* Wrote Oozie scripts and set up workflow using Apache Oozie workflow engine for managing and scheduling Hadoop jobs.
* Used Hive to analyze data ingested into HBase by using Hive-HBase integration and compute various metrics for reporting on the dashboard.
* Worked with HIVE data warehouse infrastructure-creating tables, data distribution by implementing partitioning and bucketing, writing, and optimizing the HQL queries.
* Developed PIG UDFs for manipulating the data according to Business Requirements and worked on developing custom PIG Loaders.
* Involved in loading data from UNIX file system to HDFS.
* Responsible for creating Hive tables, loading data, and writing hive queries.
* Created Hive External tables and loaded the data into tables and query data using HQL.
* Overseen importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
* Created and maintained technical documentation for launching Hadoop Clusters and for executing Hive queries and Pig Scripts.
* Extracted the data from Teradata into HDFS using Sqoop.
* Exported the patterns analyzed back to Teradata using Sqoop.
* Experience in Monitoring System Metrics and logs for any problems adding, removing, or updating Hadoop Cluster.
* Involved in scheduling Oozie workflow engine to run multiple Hives and pig jobs and used Oozie workflows for batch processing and scheduling workflows dynamically.
* Involved in requirement analysis, design, coding, and implementation phases of the project.

**Environment:** Hadoop, Spark, Scala, MapReduce, HDFS, Hive, Java, SQL, Cloudera Manager, Sqoop, Oozie, Zookeeper, Apache Pig, Python, Java, SSRS, HBase, Cassandra, PySpark, Apache Kafka, HIVE, FLUME, ETL, UDF.

**ETL Developer/ Data Engineer**

**Client Name: Honeywell CAESS, Chennai, INDIA**

**Duration: 2015 June to 2018 April**

**Responsibilities:**

* Collaborate with Lead Developers, System Analysts, Business Users, Architects, Test Analysts, Project Managers, and peer developers to analyze system requirements.
* Worked with SQL, PL/SQL procedures and functions, stored procedures, and packages within the mappings.
* Involved in all activities related to the development, implementation, and support of ETL Process using Informatica Power Center 10.x
* Worked with most of the transformations such as the Source Qualifier, Expression, Aggregator and Connected & Unconnected lookups, Filter, Router, Sequence Generator, Sorter, Joiner, SQL, and Update Strategy.
* Developed complex store procedures using input/output parameters, cursors, views, triggers, and complex queries using temporary tables and joins.
* Develop scripts for loading the data into the tables using Fast Load, Multi Load and BTEQ utilities of Teradata.
* Used control-M to schedule Jobs.
* Used Snowflake schema to be joined with the fact table.
* Involved in requirement analysis, ETL design and development for extracting data from the source systems like salesforce, Mainframe, DB2, Oracle, flat files.
* Responsible for determining the bottlenecks and fixing the bottlenecks with performance tuning.
Extensively involved in the Analysis, design, and Modeling. Worked on Snowflake Schema, Data Modeling, Data Elements, Issue/Question Resolution Logs, and Source to Target Mappings, Interface Matrix, and Design elements.
* Design and develop logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema and Slowly Changing Dimensions
* Work on Test Driven Development and conduct unit testing, system testing and user acceptance testing.
* Create deployment packages to deploy the developed Informatica mappings, Mapplets, Worklets and workflows into test and then to production.
* Trouble shoots any deployment issues and coordinates to deploy the code into production on the target date.
* Experience in using Microsoft BI studio products like SSIS, SSAS, SSRS for implementation of ETL methodology in data extraction, transformation, and loading.
* Created dashboards for analyzing POC data and applied Filter Actions between different worksheets and dashboards.
* Designed and developed various analytical reports from multiple data sources by blending data on a single worksheet in Tableau Desktop.
* Participate in agile daily stand-up, sprint planning, sprint review, retrospective, Backlog refinement and Feature Overview meetings.

**Environment:** Tableau, Informatica, ETL, Data Modelling, SQL, Oracle, DB2, Control-M, T-SQL, SSIS, SSRS, SSAS.