**P Vishnuvardhani**

**479-437-5046**

**sales21@kteksoft.com**

**PROFESSIONAL SUMMARY:**

• 11 years of experience as a Data Engineer with data applications, relational databases, NoSQL databases, data warehousing, and cloud technologies like AWS, Azure, and GCP.

• Excellent working experience in Scrum / Agile framework and Waterfall project execution methodologies.

• Experienced in using R Programming, SAS, Python, Tableau, and Power BI for data cleaning, data visualization, risk analysis, and predictive analytics.

• Experience with NoSQL databases like HBase, Cassandra, and MongoDB and SQL databases like Teradata, Oracle, PostgreSQL, and SQL Server.

• Experience using MAVEN and ANT as build tools for building deployable artifacts from source code.

• Experienced in branching, tagging, and maintaining the version across the environments using SCM tools like GIT, Subversion (SVN), and CVS on Linux and Windows platforms.

• Experienced in SQL programming skills and developed Stored Procedures, Triggers, Functions, and Packages using SQL, PL/SQL.

• Hands-on experience on Hadoop, HDFS, Hive, Sqoop, Pig, HBase, Oozie, Flume, Spark, MapReduce, Cassandra, Zookeeper, YARN, Kafka, Scala, PySpark, Airflow, Snowflake, SQL, Python.

• Strong hands-on experience with AWS services, including but not limited to EMR, S3, EC2, route S3, RDS, ELB, Dynamo DB, Glue, SNS, SQS, Cloud Formation, etc.

• Experience in Dimensional Data Modeling Star Schema, Snowflake Schema, Fact, and Dimensional Tables, concepts like Lambda Architecture, and Batch processing, Oozie.

• Good work experience with UNIX/Linux commands, scripting, and deploying the applications on the servers.

• Experience developing and scheduling ETL workflows in Hadoop using Oozie with the help of deployment and managing Hadoop cluster using Cloudera and Horton works.

• Hands-on Experience using Visualization tools like Tableau, Power BI, Grafana, Infogram, Charts, etc.

• Experience in Azure Cloud Services (PaaS & IaaS), Azure Databricks, Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure HDInsight, Key Vault, Azure Data Lake for data ingestion.

• Experienced with Spark Core, SQL, Spark MLlib, Spark GraphX, and Spark Streaming for processing and transforming complex data using in-memory computing capabilities in Scala.

• Experience building data pipelines and computing large volumes using Azure Data Factory.

• Strong experience and knowledge of real-time data analytics using Spark Streaming, Kafka, and Flume.

• Utilized big data tools for MLOps like GCP, Big Query, and DataProc for streamlining data lakes. AutoML for automating the model building process.

• Experience developing and designing scalable systems using Hadoop technologies in multiple environments.

• Experience in developing Map Reduce Programs using Hadoop to analyze big data as required.

• Good experience using Sqoop to import data into HDFS from RDBMS and vice-versa.

• Expertise and Vast knowledge of Enterprise Data Warehousing, including Data Modeling, Data Architecture, Data Integration (ETL/ELT), and Business Intelligence.

• Designed and implemented database schemas, indexes, views, and other database objects to optimize query performance and improve data accessibility by using T-SQL.

• Created Snowflake Schemas by normalizing the dimension tables as appropriate, and creating a Sub Dimension named Demographic as a subset of the Customer Dimension,

• Hands-on experience in test-driven development (TDD), Behavior-driven development (BDD), and acceptance test-driven development (ATDD) approaches.

• Experienced data pipelines using Kafka and Akka for handling large terabytes of data.

• Hands-on experience in using message brokers like RabbitMQ.

• Managed migration of on-prem servers to AWS by creating golden images for upload and deployment

• Created AWS VPC network for the installed Instances and configured security groups and Elastic IPs Accordingly.

• Developed AWS Cloud formation templates to create custom sized VPC, subnets, EC2 instances, ELB, and security groups.

• Involved in the development of real-time streaming applications using Kafka pyspark, and Apache Flink, on distributed Hadoop Clusters.

• Design and developed Flink pipelines to consume streaming data from Kafka.

• Implemented real-time streaming ingestion using Kafka and Spark Streaming

• Loaded data using Spark-streaming with Scala and Python

• Knowledge of analyzing data interactively using Apache Spark and Apache Zeppelin.

• Good Knowledge in understanding the Apache Storm-Kafka pipelines.

• Provided full life cycle support to logical/physical database design, schema management, and deployment. Adapt to the database deployment phase with strict configuration management and controlled coordination with different teams.

• Designed solutions to process high-volume data stream ingestion, processing, and data provisioning using Hadoop ecosystems Hive, Pig, Scala, and Druid.

• Familiar with the latest software development practices such as Agile Software Development, Scrum, Test Driven Development (TDD), and Continuous Integration (CI).

• Utilized analytical applications like R, SPSS, Rattle, and Python to identify trends and relationships between different pieces of data, draw appropriate conclusions and translate analytical findings into risk management and marketing strategies that drive value.

• Extensive hands-on experience in using distributed computing architectures such as AWS products (EC2, Redshift, EMR, and Elastic search), Hadoop, Python, Spark, and effective use of Azure SQL Database, Map Reduce, Hive, SQL, and PySpark to solve big data type problems.

• Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design, and implementing RDBMS-specific features.

• Experienced with CI/CD tools like Docker, Kubernetes, and Jenkins.

• Experience maintaining the big data platform using open-source technologies such as Spark and Elastic Search.

• Experience in Service Oriented Architecture using Web Services like SOAP & Restful.

• Experience converting Hive queries into Spark transformations using Spark RDDs and Scala.

• Experience working with different ETL tool environments like SSIS and Informatica and reporting tool environments like SQL Server Reporting Services (SSRS).

**TECHNICAL SKILL SET:**

|  |  |
| --- | --- |
| Hadoop Components / Big Data | HDFS, Hue, MapReduce, PIG, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, pysparkAirflow, Kafka Snowflake, |
| Languages | Scala, SQL, Python, Hive QL, KSQL. Boto3  |
| IDE Tools | Eclipse, IntelliJ, pycharm, VSCode. |
| Cloud platform | AWS, Azure |
| Reporting and ETL Tools | Tableau, Power BI, Talend, Pentaho, AWS GLUE. |
| Databases | Oracle 19, SQL Server, MySQL, Druid, MS Access, NoSQL Database (HBase, Cassandra, Mongo DB), T-SQL |
| Big Data Technologies | Hadoop, HDFS, Hive, Pig, Oozie, Sqoop, Spark, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Flink, Flume, Airflow, Informatica, Snowflake, DataBricks, Kafka, Cloudera, RabbitMQ. |
| Data Analysis Libraries:  | Pandas, NumPy, SciPy, Scikit-learn, NLTK, Plotly, Matplotlib |
| BI Tools:  | Alteryx, Tableau Power BI, Sisense, Looker. |
| Containerization | Docker, Kubernetes |
| CI/CD Tools | Jenkins, Bamboo, GitLab  |
| Operating Systems |  UNIX, LINUX, Ubuntu, CentOS. |
| Software Methodologies | Agile, Scrum, Waterfall |
| Reporting Tools  | PowerBI, Qlikview, Tableau, Crystal reports XI, Business Intelligence, SSRS, Business Objects 5. x/ 6.x, Cognos7.0/6.0. |
| Frameworks | Django, Flask, Fast API. |

**PROFESSIONAL WORK EXPERIENCE**

**Client: Michaels - Dallas, TX FEB 2023 - Present**

**Role: Lead AWS Data Engineer**

**Responsibilities:**

* Participated in the analysis, design, and development phase of the **Software Development Lifecycle (SDLC).**
* Worked on Creating a data pipeline of gathering, cleaning, and optimizing data using Hive and Spark.
* Responsible for building scalable distributed data solutions using an EMR environment with Amazon EMR.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the data got from Kafka and Persisted into HDFS.
* Designed and developed data integration programs in a Hadoop environment with NoSQL data store Cassandra for data access and analysis.
* Writing complex SQL queries, stored procedures, and triggers in DB2 for data retrieval, and manipulation purpose.
* Hands-on experience with ETL process using DB2, including data extraction and transformation.
* Worked onHadoop services such as HDFS, YARN, Pig, Hive, Hbase, Kafka, MapReduce, Sqoop, Oozie, Zookeeper, NIFI, Airflow, and involved in analyzing log data to predict the errors by using Apache Spark.
* Involved in various phases of development, analyzed and developed the system through Agile and Scrum
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python, and Scala.
* Involved in scheduling the Oozie workflow engine to run multiple Hive and Pig jobs.
* Involved in writing UNIX shell scripts and automating the ETL processes using UNIX shell scripting.
* Implemented Spark using Scala and Data frames and Spark SQL API, Data Frames, and Pair RDDs for faster processing of data and created RDDs, Data Frames, and datasets.
* Stage the API or Kafka Data (in JSON file format) into Snowflake DB by flattening the same for different functional services.
* Developed Nifi workflow to pick up the data from the rest API server, the data lake, and the SFTP server and send that to the Kafka broker.
* Worked on Apache Nifi to automate the data movement between RDBMS and HDFS.
* Worked on optimization of Hive queries using best practices and the right parameters and using technologies like Hadoop, YARN, Python, and Pyspark.
* Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and S3 data lake.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
* Implement One time Data Migration of Multistate level data from SQL server to Snowflake by using Python and SnowSQL.
* Day-to-day responsibility includes developing ETL Pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries in Snowflake.
* Designing and implementing data model using advanced SQL for storage, retrieving and manipulating the data.
* Used Spark SQL for Scala & amp, a Python interface that automatically converts RDD case classes to schema RDD.
* Created reports and dashboards in jira to project status to report to finance team.
* Worked on Dimensional Data modelling in Star and Snowflake schemas and Slowly Changing Dimensions (SCD).
* As a part of MLOps working on the migration of various pyspark models from on-perm Hadoop clusters to GCP using Dataproc and Oozie workflow.
* Using the advanced SQL to perform complex data analysis like aggregation, and statistical calculations.
* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Experience with building the streaming applications using flink and kafka stream.
* Worked on apache flink to implement the transformation on data streams for aggregating and update state.
* Creating Lambda functions with Boto3 to deregister unused AMIs in all application regions to reduce the cost of EC2 resources.
* Solutions on Snowflake Cloud, AWS Redshift, Informatica Intelligent Cloud Services (ICS - (DI) & Informatica PowerCenter integrated with multiple Relational databases (MySQL, Teradata, Oracle, SQL server, DB2)
* Designed solutions to process high-volume data stream ingestion, processing, and data provisioning using Hadoop ecosystems Hive, Pig, Scala, and Druid.
* Responsible for data services and data movement infrastructures with good experience with ETL concepts, building ETL solutions, and Data modeling.
* By using the druid powering GUIs of analytical applications, or as a backend for highly concurrent APIs that need fast aggregations.
* Extracted the data from HDFS using Hive and performed data analysis using Spark with Scala, PySpark, and Redshift for feature selection and created nonparametric models in Spark.
* Created Sqoop Scripts to import and export customer profile data from RDBMS to S3 buckets.
* Implemented continuous integration and deployment using CI/CD tools like Jenkins, GIT, and Maven.
* Worked with Spark to improve performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames, and Pair RDDs.
* working on Apache NiFi, like executing Spark script and Sqoop scripts through NiFi, creating scatter and gathering patterns in NiFi, Ingesting data from Postgres to HDFS, and Fetching Hive.
* Involved in working with ELASTIC MapReduce (EMR) and setting up environments on Amazon AWS EC2.
* Developed end-to-end data processing pipelines that begin with receiving data using distributed messaging systems, Kafka, for persisting data into Cassandra.
* Strong problem-solving and troubleshooting skills, with the ability to analyze and resolve complex database-related issues in DB2, ensuring system stability and data accuracy.
* Used Kafka and Kafka brokers, initiated the spark context, processed live streaming information with RDD, and Used Kafka to load data into HDFS and NoSQL databases.
* Documented and communicated best practices for ML operations on CP, including security, compliance, and cost optimization considerations.
* Developed ETL Specification Design document containing detailed information on ETL processing, mapping/workflow specifications, exception handling process, staging, and data warehouse schemas, etc.
* Updated Python scripts to match training data with our database stored in AWS Cloud Search so that we would be able to assign each document a response label for further classification.
* Ability to troubleshoot and debug ML Models and pipelines with experience using relevant monitoring and logging tools.
* Expertise in containerization and orchestration technologies such as Docker and Kubernetes with experience in deploying ML models.
* Involved in deploying the Big Data Hadoop application using Talend on cloud AWS (Amazon Web Services).
* Involved in Installing and configuring applications like the docker tool and Kubernetes for orchestration purposes.
* Worked on huge datasets stored in AWS S3 buckets and used spark data frames to perform preprocessing in Glue.
* Managed Docker orchestration and docker containerization using Kubernetes.
* Used Kubernetes to orchestrate docker container deployment, scaling, and management.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Experienced in developing BI dashboards and reports using pentaho reports and pentaho dashboards.
* Experience with Pentaho on AWS elastic map reduce that uses Hadoop.
* Used bunch of transformations in pentaho including row normalizer, row denormalizer, database lookup, database join, add sequences, add constants and various types of input and output for various data sources including tables, Access, and CSV files.
* Created and maintained ETL (Extract-Transform-Load) packages using T-SQL and SQL Server Integration Services (SSIS) to automate data integration processes between different systems.
* Used pentaho design studio for creating customer parameters as well as generating reports.
* Developed automation system using PowerShell scripts and JSON templates to remediate the services.
* Created data pipeline for different ingestion events, aggregating consumer response data in AWS S3 bucket into Hive external tables in HDFS location to feed Tableau dashboards.
* Worked extensively on AWS Components such as EC2, S3, RedShift, RDS, Route53, EMR, Elastic Search, IAM, VPC, KMS, Kinesis, Lambda, API Gateway, AIM, ELK, ELB, EKS, ECS, SNS, Cloud Watch, Cloud Trial, Cloud Formation, Athena, AWS Glue.
* Worked with business analysts and data scientists to understand their data needs and translate them into T-SQL scripts that produce accurate and timely results.
* Conducted data profiling and data cleansing activities to identify data quality issues and implement data cleansing rules using T-SQL scripts.
* Designed and implemented end-to-end data analytics and automation systems, integrating custom visualization tools using R and Tableau.
* Converted reports from Tableau and looker.
* Expert in building the looker dashboards and looks.
* Migrating the dashboards from other BI tools to Looker.
* Involved in identifying production bugs in the data using stack driver logs in GCP.
* Converted SAS code to python/spark-based jobs in cloud data proc/big query in GCP.
* Developed Tableau visualizations and dashboards using Tableau Desktop.
* Involved in NoSQL technologies like MongoDB and Cassandra and relational databases like Oracle, SQLite, PostgreSQL, and MySQL databases.
* Worked on ETL tool Informatica, Oracle Database and PL/SQL, Python, and Shell Scripts.

**Environment:** Agile, AWS,Cloud Watch, Cloud Trial, Cloud Formation, Athena, AWS Glue, MongoDB, Docker, ETL, GCP, snowflake, MapReduce, Sqoop, API, Oozie, Looker, Jira, Zookeeper, Pentaho, GIT, Python, Hadoop, Map Reduce, HDFS, Hive, Presto, Apache Flink, T-SQL, Python, Druid, DB2, Streaming, SQL, Amazon RDS, Amazon EC2, S3, CloudWatch, Spark, Scala, AWS, Git, Kafka, RedShift, DynamoDB, PostgreSQL

**Client: State street - Princeton, NJ Jul 2021 - DEC 2022**

**Role: Sr. Data Engineer**

**Responsibilities**

* Used the Agile and Scrum methodology to build the different software development life cycle phases.
* Improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDDs, and YARN.
* Developed ELT processes from the files from abinitio, google sheets in GCP, with compute being data prep, data proc(pyspark), and Bigquery.
* Involved in identifying production bugs in the data using stack driver logs in GCP.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators**.**
* Used Cloud shell SDK in GCP to configure the services data proc, storage, and BigQuery.
* Handled importing data from various data sources, performed transformations using Hive and Map Reduce, and loaded data into HDFS.
* Deployed applications using the Jenkins framework, integrating Git- version control.
* Developed workflow in Oozie to automate loading the data into HDFS and pre-processing, analyzing, and training the classifier using MapReduce jobs, Pig jobs, and Hive jobs.
* Worked on SQS Queue receiver using Spark Streaming context to consume the data from the extended queue and integrated with ETL Functions.
* Converted SAS code to python/spark-based jobs in cloud data proc/big query in GCP.
* Used Scala to convert Hive / SQL queries into RDD transformations in Apache Spark.
* Developed multiple POCs using Scala and deployed them on the Yarn cluster, compared Spark performance with Hive and SQL / Teradata.
* Used Cloud shell SDK in GCP to configure the services Data Proc, Storage, and BigQuery.
* Developed scripts using PySpark to push the data from GCP to the third-party vendors using their API framework.
* Wrote complex SQL queries using stored procedures, common table expressions (CTEs), and temporary tables to support Power BI reports.
* By using the druid powering use cases where real-time ingest, fast query performance, and high uptime.
* Worked on python scripts to import data from sources like MS SQL Server, SQL Lite, and Oracle DB.
* Developed and run UNIX shell scripts and implemented auto deployment process.
* Responsible for designing and implementing data models and schema designs that are optimized for Druid's columnar storage format**.**
* Worked on JIRA for defect/issues logging & tracking and documented all my work using CONFLUENCE.
* Involved on analyzed large and critical datasets using HDFS, HBase, Hive, HQL, PIG, Sqoop, and Zookeeper.
* Worked on the Design, Development, and Documentation of the ETL strategy to populate the data from the various source systems using the Talend ETL tool into the Data Warehouse.
* Worked on analyzing and examining customer behavioral data using MongoDB.
* Worked with GCP cloud using GCP Cloud storage, Data-Proc, Data Flow, Big- Query, G - Cloud function, Google Cloud Composer, Cloud dataflow, Pub/Sub cloud shell, GSUTIL, BQ Command line utilities, Data Proc, Stack driver.
* Proficient with container systems like docker and container orchestration, Kubernetes worked with terraform.
* Experience with container-based deployment using Docker, and Kubernetes.
* Worked with the openshift platform in managing the docker container and Kubernetes clusters.
* Utilized Kubernetes for the run time environment of the CI/CD to build, test and deploy.
* Implemented Data Validation using MapReduce programs to remove unnecessary records before moving data into Hive tables.
* Involved in using Tomcat Apache servers and Docker containers for deployment.
* Involved in working with some testing tools like Bugzilla and JIRA.
* Imported data from relational data sources like Oracle and Teradata to HDFS using Sqoop.
* Designed, built, and coordinate an automated build & release CI/CD process using Gitlab, Jenkins, and Puppet on hybrid IT infrastructure.
* Mastered the ability to design and deploy rich Graphic visualizations using Tableau and Converted existing Business object reports into tableau dashboards.
* Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring, and cloud deployment manager.
* Used Spark API over Cloudera Hadoop YARN to analyze Hive’s data.

**Environment:** Agile, Azure, Hadoop, MapReduce, Oozie, Pig, Zookeeper, HDFS, Hive, Docker, Druid, GIT, ETL, Python, JIRA, NoSQL, Mongo DB, Kubernetes, Cassandra, Tableau, Power BI.

**Client: Discovery Inc, Charlotte, NC Jan 2020 – Jun 2021**

**Sr. Data Engineer**

**Responsibilities:**

* Involved in working with Agile methodology, including test-driven and pair-programming concepts.
* Involved in implementing nine node CDH4 Hadoop cluster on red hat LINUX.
* Involved in importing data from various sources, performed transformations using Hive and Map Reduce, and loaded data into HDFS.
* Developed a data pipeline using MapReduce, Flume, Sqoop, and Pig to analyze customer data into HDFS.
* Acquired an excellent understanding and worked on NoSQL databases such as HBase and Cassandra.
* Involved in scheduled Oozie workflow engine to run multiple Hive and Pig jobs, which independently run with time and data availability.
* Involved in working with Hadoop technologies like MapReduce, HDFS, HBase, Zookeeper, Hive, Pig, Sqoop, Spark, Pyspark, Oozie, Storm, Yarn, Snowflake, Flume, Nifi, and Airflow.
* Involved in loading and transforming large sets of structured, semi-structured, and unstructured data Hadoop concepts.
* Implemented a distributing messaging queue to integrate with Cassandra using Apache Kafka.
* Involved in using AWS utilities such as EMR, S3, and Cloud Watch to run and monitor Hadoop and Spark jobs on AWS.
* Designed and developed APIs to share data with cross-functional teams using Fast API frameworks.
* Expertise in developing restful microservices using python and fast API frameworks.
* Developed web applications using FastAPI to deliver highly performant and scalable RESTful APIs.
* Implemented authentication and authorization mechanisms in FastAPI to secure endpoints.
* Involved in working on analysis tools like Tableau, and Splunk for regression analysis, pie charts, and bar graphs.
* Developed views and templates with Python and Django's view controller and template language to create a user-friendly website interface.
* Created entire application using Python, Django, MySQL and Linux.
* Developed complex MapReduce streaming jobs using Java language implemented Using Hive and Pig and MapReduce Programs using Java to perform various ETL, cleaning, and scrubbing tasks.
* Developed spark applications in python (PySpark) on the distributed environment to load the massive number of CSV files with different schema into Hive ORC tables.
* Rewrite existing Python/Django modules to deliver certain format of data.
* Used Django Database API's to access database objects.
* Involved in ETL architecture enhancements to increase the performance using query optimizer.
* Involved in working with some AWS cloud technologies like Cloud Watch, Cloud Trial, Cloud Formation, Docker, Kubernetes, Terraform, Athena, AWS Glue, and AWS Sage Maker.
* Worked on querying multiple databases like Snowflake, Netezza, UDB, and MySQL for data processing.
* Involved in working with Continuous Integration and Deployment (CI/CD) using Jenkins and Docker.
* Worked with AWS, where the cluster was built using EC2 instances, stored data in S3, and used Athena Serverless Query Services.
* Worked on Python Open stack API and used Python scripts to update content in the database and manipulate files.
* Automated nightly build to run quality control using Python with BOTO3 library to make sure the pipeline does not fail which reduces the effort by 70%.
* Involved with the development of Ansible playbooks with Python and SSH as wrappers for the management of AWS node configurations and testing playbooks on AWS instances.
* Developed Python AWS serverless lambda with concurrent and multi-threading to make the process faster and asynchronously execute the callable.
* Chunking the data to convert them from larger data sets to smaller chunks using python scripts which will be useful for faster data processing.
* Responsible for monitoring and performance of Druid clusters and performing routine maintenance tasks such as data backups and system upgrades.
* Responsible for integrating Druid with other systems within an organization, such as business intelligence tools, data warehouses, and other data processing systems.
* Responsible for optimizing queries to ensure that they execute as quickly as possible and return accurate results.
* Created stage based DW supported by that’s completely implemented in pentaho kettle.
* Used pentaho import export utility to migrate pentaho transformations and jobs from one environment to other.
* Used pentaho enterprise console to monitor the ETL jobs on production database.
* Worked with the data from Mainframe DB2 and ADABAS using Apache Sqoop and stored it in AWS S3 to create the data lake for Apache Spark, Athena, and Redshift
* Experience in setting up the data actions in looker.
* Experience in looker custom table calculations like offset etc.
* Setting up the data actions in looker table calculations in looker.
* Wrote Pig Scripts to generate Map Reduce jobs and performed ETL procedures on the data in HDFS.
* Exploring Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, Datasets, RDDs, and Spark YARN.
* Kept up-to-date with the latest T-SQL and SQL Server features and best practices to improve development efficiency and quality.
* Developed complex T-SQL queries and stored procedures to extract, transform and load data from various sources into the data warehouse.
* Implement the function to send and receive messages from RabbitMQ synchronously or asynchronously.
* Developed Spark applications using Scala and Python and implemented Apache Spark for data processing from various streaming sources.
* Developed ETL jobs and pipelines using tools such as IBM DataStage and Apache Nifi.
* Involved in writing and testing Python scripts to create new data files for Linux server configuration using a Python template tool.
* Conducted data profiling and data cleansing activities to identify data quality issues and implement data cleansing rules using T-SQL scripts.
* Involved in converting Cassandra/Hive/SQL queries into Spark using Spark RDDs and Scala Python.
* Involved in checking the data and table structure in the PostgreSQL & Redshift databases and running the queries to generate reports.
* Developed SQL scripts for data loading and table creation in Teradata and Hadoop, HIVE, Oozie, and Sqoop.
* Worked with Maven to build scripts and set up the Log4J Logging framework.

**Environment:** Spark, YARN, HIVE, Pig, Scala, Druid, Pentaho, NiFi, TDD, Python, T-SQL, AWS, Hadoop, Fast API, Azure, Dynamo DB, NOSQL, Sqoop, RabbitMQ, MYSQL, Looker.

**Client: Lowes Corporate - Charlotte, NC DEC 2018 - DEC 2019**

**Role: Azure Data Engineer**

**Responsibilities:**

* Used Oozie workflow engine to manage interdependent Hadoop jobs and automate several Hadoop jobs such as Java MapReduce Hive, Pig, and Sqoop.
* Ability to collaborate with cross-functional teams such as data scientists, software engineers, and business stakeholders to drive ML model deployment and monitoring in Cloud.
* Involved in working with the Agile and Scrum Methodologies.
* Analyzed critical datasets using Cloudera, HDFS, HBase, MapReduce, Hive, Pig, Sqoop, Spark, and Zookeeper.
* Worked on MongoDB NoSQL data modeling, tuning, disaster recovery, and backup.
* Used NoSQL database Hbase and created Hbase tables to load large sets of semi-structured data from various sources.
* Solution was designed and developed using java, Spring.
* Design and developed services to persist and read data from Hadoop, HDFS, hive and writing java-based MapReduce batch.
* Wrote Hive and Pig scripts as an ETL tool to perform transformations, event joins, filter traffic, and some pre-aggregations before storing them in the HDFS.
* Developed data pipeline using Flume, Sqoop, Pig, and MapReduce to ingest customer behavioural data.
* Developed Spark code using Scala and Spark-SQL for faster testing and processing of data.
* Used Spark API over Cloudera Hadoop YARN to analyse Hive’s data.
* Involved in moving all log files generated from various sources to HDFS for further processing through Flume.
* Developed Java code to generate, compare & merge AVRO schema files.
* Worked on reading and writing multiple data formats like JSON, ORC, and Parquet on HDFS using PySpark.
* Prepared the validation report queries, executed after every ETL runs, and shared the resultant values with business users in different project phases.
* Involved in installing and configuring Pentaho BI server for ETL and reporting services.
* Worked with cloud-based technology like Redshift, S3, AWS, EC2 Machine, etc., extracting the data from the Oracle financials and the Redshift database, Creating Glue jobs in AWS, and loading incremental data to the S3 staging area and persistence area.
* Involved in Functional Testing, Integration testing, Regression Testing, Smoke testing, and performance Testing and tested Hadoop Map Reduce developed in python, pig, and Hive.
* Designed and implemented data models, routes, and controllers in FastAPI using Python.
* Worked with FastAPI plugins and extensions such as SQLAlchemy and Pydantic to streamline development and improve code maintainability.
* Developed a common Flink module for serializing and deserializing AVRO data by applying Schema.
* Experience in developing enterprise-level solutions using batch processing and streaming framework using Kafka, and Apache Flink.
* Integrated FastAPI with various databases including PostgreSQL, MongoDB, and MySQL to persist data.
* Collaborated with cross-functional teams to design, develop, and deploy complex APIs using FastAPI while adhering to best practices and project requirements.
* Deployed FastAPI applications using Docker and Kubernetes to containerize and manage services at scale.
* Used Hive to analyse the partitioned and bucketed data, computed various metrics for reporting & used the hive optimization techniques during joins and best practices in writing hive scripts using HiveQL.
* Importing and exporting data into HDFS and Hive using Sqoop. Writing the HIVE queries to extract the data.
* Worked to design a Spark model for the existing MapReduce model and Migrated MapReduce models to Spark Models using Scala.
* Implemented the Machine learning algorithms using Spark with Python.
* Experience in CI/CD principles and experience building and deploying ML models using tools such as Jenkins and GIT.
* Experience with performance optimization and scaling of ML pipelines, with the ability to identify in distributed systems.
* Working on cloud environment such as AWS MLOPS (Sagemaker, ECR, Codebulid, Docker, and Kubernetes) and cloud functions, Terraform, and git CI/CD.
* Implemented Spark using Scala and SparkCore, Spark Streaming, and SparkSQL API for faster data processing instead of MapReduce in Java.
* Created monitors, alarms, notifications, and logs for Lambda functions, Glue Jobs, and EC2 hosts using CloudWatch and used AWS Glue for the data transformation, validation, and data cleansing.
* Create a Pyspark frame to bring data from DB2 to Amazon S3.
* Handled importing of data from various data sources, performed transformations using Hive, and MapReduce, loaded data into HDFS, and Extracted the data from MySQL into HDFS using Sqoop.
* Integrated Apache Storm with Kafka to perform web analytics and click stream data from Kafka to HDFS.

**Environment:** Agile, AWS, S3,EC2, S3, RedShift, RDS, Route53, EMR, Elastic Search, Cloudera, HDFS, HBase, MapReduce, Hive, Pig, Sqoop, Flink, Spark and Zookeeper, Fast API, Mongo DB, Cassandra, GIT, Spark SQL, Spark RDD, Python, Tableau.

**Client: AG FIRST FARMERS CREDIT BANK - Columbia, SC Nov 2016 - Nov 2018**

**Role: Data Engineer**

**Responsibilities:**

* Leveraged Sqoop to seamlessly import and export data from Oracle and PostgreSQL into HDFS, facilitating data analysis and processing tasks.
* Successfully migrated existing MapReduce programs to Spark models using Python, enhancing performance and scalability of data processing workflows.
* Spearheaded the migration of data from Data Lake (Hive) into S3 Bucket, ensuring efficient data storage and accessibility in cloud environments.
* Conducted thorough data validation between data residing in Data Lake and S3 bucket, ensuring data consistency and integrity across platforms.
* Utilized Spark DataFrame API on the Cloudera platform to perform advanced analytics on Hive data, extracting valuable insights and driving informed decision-making.
* Designed batch processing jobs using Apache Spark, resulting in a ten-fold increase in processing speeds compared to traditional MapReduce jobs.
* Employed Qualtrics and developed sentiment scores using NB classifier, calculating Net Promoter score, and administering PostgreSQL Database and NoSQL Data warehouse.
* Integrated MSSQL with data modeling systems such as HDFS, S3, or Azure Data Lake Storage, facilitating effective data mapping and integration.
* Implemented Kafka for real-time data ingestion, creating different topics for reading and processing data streams efficiently.
* Orchestrated the movement of data from S3 bucket to Redshift data warehouse for generating reports, ensuring seamless data transfer and availability for analytics.
* Developed Hive queries for data analysis, addressing business requirements and extracting actionable insights from structured data sources.
* Executed the migration of an existing on-premises application to AWS, leveraging cloud infrastructure for enhanced scalability and reliability.
* Developed PIG Latin scripts to extract data from web server output files and load it into HDFS, streamlining data ingestion processes.
* Utilized Hive to analyze partitioned and bucketed data, computing various metrics for reporting and decision-making purposes.
* Developed numerous Spark UDFs and UDAFs in Hive to address specific functional requirements not supported by existing Hive and Spark SQL functionalities.
* Developed Spark scripts to import large files from Amazon S3 buckets, optimizing data transfer processes for efficiency and scalability.
* Implemented SNS notifications and CloudWatch events to monitor and alert data pipeline failures, ensuring timely response to potential issues.
* Established external Hive tables on top of S3 datasets in Parquet format, enabling end-users to query data efficiently and analyzing reported data issues.
* Designed and implemented complex applications and distributed systems into public cloud infrastructure (AWS, GCP, Azure), ensuring scalability and reliability of data solutions.
* Led the migration of on-premises data to Azure Data Lake and Stored (ADLS) using Azure Data Factory, streamlining data storage and accessibility in cloud environments.
* Managed Hive data model changes and utilized AWS Glue as Hive Metastore for Hive external tables, ensuring data consistency and accessibility.
* Loaded data from AWS S3 to Redshift, facilitating data analytics and reporting in cloud environments.
* Created S3 event-based AWS Lambda functions to load data into PostgreSQL AWS RDS service, automating data loading processes for efficiency.
* Developed ad-hoc AWS Lambdas using Python to support production activities, enhancing automation and efficiency of data operations.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs and Scala, optimizing data processing workflows for performance and scalability.
* Implemented various performance optimization techniques such as distributed cache, partitioning, and bucketing in Hive, enhancing query performance and resource utilization.
* Demonstrated expertise in Spark platform parameters like memory, cores, and executors, optimizing Spark job performance and resource allocation.
* Implemented Zookeeper for concurrent access to Hive tables with shared and exclusive locking, ensuring data consistency and integrity in distributed environments.

**Client: AbbVie - India** **Aug 2015 - July 2016**

**Role: Hadoop Engineer**

**Responsibilities:**

* Participated in all phases of the Software Development Life Cycle (SDLC), encompassing development, testing, implementation, and maintenance, ensuring comprehensive involvement and contribution throughout the project lifecycle.
* Installed and configured Hadoop MapReduce, HDFS, and developed multiple MapReduce jobs in Java for data cleansing and preprocessing, enabling efficient data processing and preparation tasks.
* Played a key role in loading data from the UNIX file system to HDFS, ensuring seamless data ingestion and storage in the Hadoop ecosystem.
* Installed and configured Hive and developed Hive User Defined Functions (UDFs), facilitating advanced data querying and manipulation capabilities within the Hadoop environment.
* Exported data to Azure Data Lake Stores and stored it in native formats using different sources, including relational and semi-structured data, leveraging Azure Data Factory for efficient data integration and storage.
* Managed the importing and exporting of data into HDFS and Hive using Sqoop, ensuring seamless data transfer between Hadoop and external data systems.
* Took responsibility for cluster maintenance, including adding and removing cluster nodes, cluster monitoring and troubleshooting, as well as managing and reviewing data backups and Hadoop log files.
* Handled the importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS, ensuring data integrity and accuracy throughout the process.
* Extracted data from Teradata into HDFS using Sqoop, facilitating efficient data transfer and integration from external data systems to the Hadoop environment.
* Led the migration of ETL processes from Oracle to Hive, streamlining data manipulation tasks and optimizing data processing workflows.
* Conducted optimization on Pig scripts and Hive queries to enhance efficiency and introduce new features to existing code, ensuring optimal performance and scalability.
* Worked on creating tabular models on Azure Analysis Services to meet business reporting requirements, enabling efficient data analysis and visualization.
* Utilized Hive to create Hive tables, load data, and develop Hive UDFs, facilitating flexible data querying and manipulation within the Hadoop ecosystem.
* Demonstrated expertise in Data Warehousing and Data Mining, utilizing NoSQL Databases such as HBase, Cassandra, and MongoDB to manage and analyze large datasets efficiently.
* Leveraged Sqoop to import data into HDFS and Hive from other data systems, enabling seamless data integration and transfer within the Hadoop ecosystem.
* Analyzed data by performing Hive queries and running Pig scripts to gain insights into user behavior, facilitating informed decision-making and business strategy development.
* Continuously monitored and managed the Hadoop cluster through Cloudera Manager, ensuring optimal performance, reliability, and scalability of the Hadoop environment.
* Installed Oozie workflow engine to orchestrate multiple Hive jobs, streamlining workflow management and enhancing job scheduling capabilities.
* Developed Hive queries to process data and generate data cubes for visualization, enabling efficient data analysis and reporting.

**Client: NETAPP-INC - Bengaluru, INDIA July 2012 - Jun 2015**

**Role: Data Engineer**

**Responsibilities:**

* Documented the complete process flow to meticulously outline program development, logic, testing, and implementation, ensuring comprehensive understanding and facilitating seamless application integration and coding.
* Recommended structural changes and enhancements to systems and databases, based on thorough analysis and evaluation of existing architectures and requirements.
* Conducted design reviews and technical reviews with other project stakeholders, fostering collaboration and ensuring alignment with project objectives and technical standards.
* Engaged in the complete life cycle of the project from requirements gathering to production support, actively contributing at every stage to ensure project success and continuity.
* Created comprehensive test plan documents for all back-end database modules, meticulously outlining testing strategies and procedures to ensure robustness and reliability of database functionalities.
* Leveraged MS Excel, MS Access, and SQL proficiently to write and execute various queries, facilitating data analysis and extraction tasks efficiently.
* Worked extensively on creating tables, views, and SQL queries in MS SQL Server, ensuring optimized database structures and efficient data retrieval processes.
* Collaborated with internal architects to assist in the development of current and target state data architectures, contributing insights and recommendations to enhance data management strategies and frameworks.
* Coordinated with business users to understand reporting needs and designed effective solutions based on user requirements and existing system functionalities, ensuring alignment with business objectives.
* Maintained knowledge of all areas of business operations to identify systems needs and requirements accurately, facilitating the development of tailored data solutions to meet business challenges.
* Installed and configured Apache Hadoop to test the maintenance of log files in the Hadoop cluster, ensuring smooth operation and performance of the Hadoop ecosystem.
* Installed and configured Hive, Pig, Sqoop, and Oozie on the Hadoop cluster, enabling seamless integration and operation of essential Hadoop components for data processing and workflow orchestration.
* Demonstrated proficiency in SQL programming and the creation of relational database models, ensuring the integrity and efficiency of database operations.
* Installed Oozie Workflow engine to orchestrate multiple Hive and Pig jobs, streamlining workflow management and enhancing job scheduling capabilities.
* Developed Spark scripts using Scala as per project requirements, leveraging the Spark 1.5 framework to implement efficient data processing solutions.
* Analyzed large datasets to determine optimal aggregation and reporting strategies, leveraging insights to enhance data visualization and decision-making processes.
* Responsible for building scalable distributed data solutions using Hadoop, ensuring the reliability, scalability, and performance of data processing workflows.
* Developed SQL queries, stored procedures, triggers, and functions for MySQL databases, ensuring the integrity and efficiency of database operations.