**Pavan**

**PROFESSIONAL SUMMARY:**

* Around 8 years IT experience in Analysis, Design, Development and Big Data in Scala, Pyspark, Hadoop, and HDFS environment and experience in Python.
* Implemented Big Data solutions using Hadoop technology stack, including Pyspark, Hive, Sqoop, Avro and Thrift.
* Proficiency in developing SQL with various relational databases like Oracle, SQL Server for Support of Data Warehousing and Data Integration Solutions.
* Strong experience in migrating other databases to Snowflake. In-depth knowledge of Snowflake Database, Schema and Table structures.
* Hands on Experience with dimensional modeling using star schema and snowflake models.
* Firm understanding of Hadoop architecture and various components including HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce programming.
* Involved in setting up Jenkins Master and multiple slaves for the entire team as a CI tool as part of Continuous development and deployment process
* Installed and configured apache Airflow for workflow management and created workflows in python, created the DAG’s using Airflow to run jobs sequentially and parallelly.
* Experienced in Optimizing the Pyspark jobs to run on Kubernetes Cluster for faster data processing
* Involved in converting Hive Queries into various Spark Actions and Transformations by Creating RDD and Data frame from the required files in HDFS.
* Experience in providing support to data analyst in running Hive queries and building an ETL.
* Performed Importing and exporting data into HDFS and Hive using Sqoop.
* Worked on AWS Cloud – used S3 buckets for storing the data, EMR – for spinning up the cluster and running the spark jobs, Athena – for creating external tables, IAM – for authentication, AWS GLUE – to get data from sources.
* Experienced in Designing, Architecting, and implementing scalable cloud-based web applications using AWS and Azure.
* Experienced on Migrating SQL database to Azure data Lake storage, Azure Data Factory (ADF), Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and controlling and granting database access and migrating on premise databases to Azure Data Lake store using Azure Data factory.
* Involved in Software development, Data warehousing and Analytics and Data engineering projects using Hadoop, MapReduce, Hive, and other open-source tools/technologies.
* Worked on reading and writing multiple data formats like JSON, ORC, Parquet on HDFS using PySpark.
* Experienced in requirement analysis, application development, application migration and maintenance using Software Development Lifecycle (SDLC) and Python technologies.
* Defined user stories and driving the agile board in JIRA during project execution, participate in sprint demo and retrospective.
* Strong working experience with SQL and NoSQL databases, data modeling and data pipelines. Involved in end-to-end development and automation of ETL pipelines using SQL and Python.
* Managed multiple tasks and worked under tight deadlines and in fast pace environment.
* Excellent analytical, communication skills which helps to understand the business logics and develop a good relation between stakeholders and team members.
* Good communication skills, work ethics and the ability to work in a team efficiently with good leadership skills.Implementation and Support using Agile and Waterfall Methodologies.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Big Data Ecosystem: | HDFS, MapReduce, HBase, Pig, Hive, Sqoop, Kafka Flume, Cassandra, Impala, Oozie, Zookeeper, MapR, Amazon Web Services (AWS), EMR |
| Machine Learning: | Classification Algorithms Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Gradient Boosting Classifier, Extreme Gradient Boosting Classifier, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes Classifier, Extra Trees Classifier, Stochastic Gradient Descent, etc. |
| Cloud Technologies: | AWS, Azure, Google cloud platform (GCP) |
| IDE’sIntelliJ: | Eclipse, Spyder, Jupyter |
| Ensemble and Stacking | Averaged Ensembles Weighted Averaging, Base Learning, Meta Learning, Majority Voting, Stacked Ensemble, Auto ML - Scikit-Learn, ML jar, etc. |
| Databases: | Oracle 11g/10g/9i, MySQL, DB2, MS SQL Server, HBASE |
| Programming: | Query Languages Java, SQL, Python Programming (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib, NLTK), NoSQL, PySpark, PySpark SQL, SAS, R Programming (Caret, Glmnet, XGBoost, rpart, ggplot2, sqldf), RStudio, PL/SQL, Linux shell scripts, Scala.  f |
| Data Engineer: | Big Data Tools / Cloud / Visualization / Other Tools Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, Mahout, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Salesforce, GCP, Google Shell, Linux, PuTTY, Bash Shell, Unix, etc., Tableau, Power BI, SAS, We Intelligence, Crystal Reports, Dashboard Design. |

**PROFESSIONAL EXPERIENCE**

**JPMC (Remote) May 2021 to Present**

**AWS Data Engineer.**

**Responsibilities:**

* Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services (AWS) on EC2.
* Handled AWS Management Tools as Cloud watch and Cloud Trail.
* Stored teh log files in AWS S3. Used versioning in S3 buckets where teh highly sensitive information is stored.
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Working experience with data streaming process with Kafka, Apache Spark, Hive.
* Analysed the SQL scripts and designed the solution to implement using Scala.
* Designed both 3NF data models for OLTP systems and dimensional data models using star and snowflake Schemas.
* Optimized the PySpark jobs to run on Kubernetes Cluster for faster data processing.
* Developed parallel reports using SQL and Python to validate the daily, monthly, and quarterly reports.
* Designed and Developed Real time Stream processing Application using Spark, Kafka, Scala, and Hive to perform Streaming ETL and apply Machine Learning.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Primarily Responsible for converting Manual Report system to fully automated CI/CD Data Pipeline that ingest data from different Marketing platform to AWS S3 data lake.
* Designed and developed AWS architecture, Cloud migration, AWS EMR, DynamoDB, Redshift and event processing using lambda function
* Conducted ETL Data Integration, Cleansing, and Transformations using AWS glue Spark script.
* Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline which can be written to Glue Catalog and can be queried from Athena.
* Developed an aws lambda functions to trigger the unit step functions to process the scheduled EMR job.
* Created Partitioned and Bucketed Hive tables in Parquet File Formats with Snappy compression and then loaded data into Parquet hive tables from Avro hive tables.
* Involved in designing and developing tables in HBase and storing aggregated data from Hive Table.
* Analysed the sql scripts and designed it by using PySpark SQL for faster performance.
* Developed spark applications in python (PySpark) on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
* Used Apache Spark Data frames, Spark-SQL, Spark MLlib extensively and developing and designing POC's using Scala, Spark SQL and MLlib libraries.
* Created PySpark code that uses Spark SQL to generate data frames from avro formatted raw layer and writes them to data service layer internal tables as orc format.
* Used SQL Server Integrations Services (SSIS) for extraction, transformation, and loading data into target system from multiple sources.
* Developed Airflow DAGs in python by importing the Airflow libraries.

**Environment**: AWS, JMeter, Kafka, Ansible, Jenkins, Docker, Maven, Linux, Red Hat, GIT, Cloud Watch, Python, Shell Scripting, Golang, Web Sphere, Splunk, Tomcat, Soap UI, Kubernetes, Terraform, PowerShell.

**Humana Health, Louisville, KY Feb 2020 to April 2021**

**AWS Data Engineer**

**Responsibilities:**

* Primarily Responsible for converting Manual Report system to fully automated CI/CD Data Pipeline that ingest data from different Marketing platform to AWS S3 data lake.
* Experience in developing scalable & secure data pipelines for large datasets.
* Migrated and developed the daily running jobs from on-premises to EMR spark, Scala.
* Used Data Frame API in Scala for converting the distributed collection of data organized into named columns, developing predictive analytic using Apache Spark Scala APIs
* Developed Hive queries to pre-process the data required for running the business process.
* Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline which can be written to Glue Catalog and can be queried from Athena.
* Worked on terraform scripts to deploy the resources.
* Compared the performance of the Hadoop based system to the existing processes used for preparing the data for analysis
* Developed an aws lambda functions to trigger the unit step functions to process the scheduled EMR job.
* Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as DynamoDB and MongoDB.
* Collected data using Spark Streaming from AWS S3 bucket 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.
* Designed and developed AWS architecture, Cloud migration, AWS EMR, DynamoDB, Redshift and event processing using lambda function
* Conducted ETL Data Integration, Cleansing, and Transformations using AWS glue Spark script.
* Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
* Designing and Developing Apache NiFi jobs to get the files from transaction systems into data lake raw zone.
* Very good understanding of Partitions, bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Used Spark Streaming to divide streaming data into batches as an input to spark engine for batch processing.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's.
* Created action filters, parameters and calculated sets for preparing dashboards and worksheets using PowerBI.
* Developed visualizations and dashboards using PowerBI
* Sticking to ANSI SQL language specification wherever possible, and providing context about similar functionality in other industry-standard engines (e.g. referencing PostgreSQL function documentation)
* Used ETL to implement teh Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
* Performing ETL testing activities like running teh Jobs, Extracting teh data using necessary queries from database transform, and upload into teh Data warehouse servers.
* Created dashboards for analyzing POS data using Power BI.

**Environment**: MS SQL Server 2016, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel (creating formulas, pivot tables, Hlookup, Vlookup, Macros), Spark, Python, ETL, Power BI, Tableau, Presto, Hive/Hadoop, Snowflakes, Power BI, AWS Data Pipeline, IBM Cognos 10.1, Data Stage, Cognos Report Studio 10.1, Cognos 8 & 10 BI, Cognos Connection, Cognos office Connection, Cognos 8.2/3/4, Data stage and Quality Stage 7.5

**USAA San Antonio, Texas Jan 2019 to Feb 2020**

**Senior Data Engineer.**

**Responsibilities:**

* Processed teh Web server logs by developing Multi-hop flume agents by using Avro Sink and loaded into MongoDB for further analysis, also extracted files from MongoDB through Flume and processed.
* Developed a data platform from scratch and took part in requirement gathering and analysis phase of the project in documenting the business requirements.
* Worked in designing tables in Hive, MYSQL using SQOOP and processing data like importing and exporting of databases to the HDFS, involved in processing large datasets of different forms including structured, semi-structured and unstructured data.
* Developed, built, and deployed applications using Azure stack development kit.
* Building and Installing servers through ARM Templates for Azure Portal.
* Worked on Ansible, used YAML packages for installing, configuring push to make changes in time
* Created the containers using Docker
* Built database model, Views and API's using Python for interactive web-based solutions.
* Used Python scripts to update the content in database and manipulate files.
* Generated Python Django forms to maintain the record of online users
* Developed rest API's using python with flask and Django framework and done the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Worked with Hadoop architecture and the daemons of Hadoop including Name-Node, Data Node, Job Tracker, Task Tracker, and Resource Manager.
* Used AWS data pipeline for Data Extraction, Transformation and Loading from homogeneous or heterogeneous data sources and built various graphs for business decision-making using Python matplot library
* Developed scripts to load data to hive from HDFS and involved in ingesting data into Data Warehouse using various data loading techniques.
* Written new spark jobs in Scala to analyse the data of the customers and sales history.
* Scheduled Jobs using crontab, run deck and control-m.
* Build Cassandra queries for performing various CRUD operations like create, update, read and delete, also used Bootstrap as a mechanism to manage and organize the html page layout
* Developed entire frontend and backend modules using Python on Django Web Framework and created User Interface (UI) using JavaScript, bootstrap, Cassandra with MySQL and HTML5/CSS.
* Importing and exporting data jobs, to perform operations like copying data from HDFS and to HDFS using Sqoop and developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Analyzed SQL scripts and designed the solutions to implement using PySpark.
* Used JSON and XML SerDe's for serialization and de-serialization to load JSON and XML data into Hive tables.
* Used SparkSQL to load JSON data and create Schema RDD and loaded it into Hive Tables and handled structured data using SparkSQL.
* Designed and Developed Real time Stream processing Application using Spark, Kafka, Scala, and Hive to perform Streaming ETL and apply Machine Learning.
* Developing data processing tasks using PySpark such as reading data from external sources, merge data, perform data enrichment and load in to target data destinations.
* Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
* Worked in development of applications especially in LINUX environment and familiar with all its commands and worked on Jenkins continuous integration tool for deployment of project and deployed the project into Jenkins using GIT version control system
* Managed the imported data from different data sources, performed transformation using Hive, Pig and Map- Reduce and loaded data in HDFS.
* Executed Oozie workflow engine to run multiple Hive and Pig jobs, which run independently with time and data availability and developed Oozie workflow to run job onto data availability of transactions.
* To achieve Continuous Delivery goal on high scalable environment, used Docker coupled with load-balancing tool Nginx.
* Used MongoDB to stored data in JSON format and developed and tested many features of dashboard using Python, Bootstrap, CSS, and JavaScript.
* Implemented Real time analytics on Cassandra data using thrift API.
* Designed Columnar families in Cassandra and Ingested data from RDBMS, performed transformations and exported teh data to Cassandra.
* Leading teh testing efforts in support of projects/programs across a large landscape of technologies (Unix, Angular JS, AWS, Sause LABS, Cucumber JVM, Mongo DB, GitHub, Bitbucket, SQL, NoSQL database, API, Java, Jenkins

**Environment**: Hadoop(HDFS,MapReduce),Databricks,Spark,Talend,Impala,Hive,postgresql,Jenkins,Nifi,Scala,Mongo DB, Cassandra, Python, Pig, Sqoop, Hibernate, spring, Oozie, AWS Services EC2, S3, Autoscaling, Scala, Azure, Elastic Search, DynamoDB, UNIX Shell Scripting, TEZ.

**Dhruvsoft Services Private Limited HYD, IN. September 2016 to July 2018**

**Data Engineer**

**Responsibilities:**

* Gathering data and business requirements from end users and management. Designed and built data solutions to migrate existing source data in Data Warehouse to Atlas Data Lake (Big Data)
* Analyzed huge volumes of data Devised simple and complex HIVE, SQL scripts to validate Dataflow in various applications. Performed Cognos report validation. Made use of MHUB for validating Data Profiling & Data Lineage.
* Experience in Big Data Analytics and design in Hadoop ecosystem using MapReduce Programming, Spark, Hive, Pig, Sqoop, HBase, Oozie, Impala, Kafka
* Performing hive tuning techniques like partitioning and bucketing and memory optimization.
* Worked on different file formats like parquet, orc, json and text files.
* Experience in developing scalable & secure data pipelines for large datasets.
* Partnered with ETL developers to ensure that data is well cleaned, and the data warehouse is up to date for reporting purpose by Pig.
* Supported data quality management by implementing proper data quality checks in data pipelines.
* Involved in the data support team as role of bug fixes, schedule change, memory tuning, schema changes loading the historic data.
* Build machine learning models to showcase big data capabilities using Pyspark and MLlib.
* Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
* Implemented data streaming capability using Kafka and Talend for multiple data sources.
* Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
* Used Oozie workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as Java map-reduce Hive, Pig, and Sqoop.
* Used spark SQL to load data and created schema RDD on top of that which loads into hive tables and handled structured using spark SQL.
* Implemented a CI/ CD pipeline with Docker, Jenkins, and GitHub by virtualizing the servers using Docker for the Dev and Test environments by achieving needs through configuring automation using Containerization.
* Involved in the development of agile, iterative, and proven data modeling patterns that provide flexibility.
* Performed data validation which does the record wise counts between the source and destination.
* Involved in the data support team as role of bug fixes, schedule change, memory tuning, schema changes loading the historic data.
* Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
* Worked on analyzing and resolving the production job failures in several scenarios.
* Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Devised PL/SQL Stored Procedures, Functions, Triggers, Views and packages. Made use of Indexing, Aggregation and Materialized views to optimize query performance.
* Hands of experience in GCP, Big Query, GCS bucket, G - cloud function, cloud data flow, Pub/sub cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver
* Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
* Worked on confluence and Jira skilled in data visualization like Matplotlib and seaborn library
* Experience implementing machine learning back-end pipeline wif Pandas, Numpy

**Environment**: Hive, AWS, Hadoop, HDFS, Python, PL/SQL, SQL, Python, R Programming, Apache Airflow, Numpy, Pandas, Jira, PIG, Tableau, Spark, Linux, Pandas, Numpy.

**CSC, Hyderabad, India June 2015 to August 2016**

**Hadoop Engineer**

**Responsibilities**

* Involved in complete Implementation lifecycle, specialized in writing custom MapReduce, and Hive
* Extensively used Hive/HQL or Hive queries to query or search for a string in Hive tables in HDFS
* Continuous monitoring and managing teh Hadoop cluster using Cloudera Manager
* Implemented Spark using Python and Spark SQL for faster processing of data
* Involved in designing physical and logical data model using ERwin Data modeling tool.
* Designed the relational data model for operational data store and staging areas, Designed Dimension & Fact tables for data marts.
* Extensively used Erwin data modeler to design Logical/Physical Data Models, relational database design.
* Created Stored Procedures, Database Triggers, Functions and Packages to manipulate the database and to
* apply the business logic according to the user's specifications.
* Created Triggers, Views, Synonyms and Roles to maintain integrity plan and database security.
* Creation of database links to connect to the other server and Access the required info.
* Integrity constraints, database triggers and indexes were planned and created to maintain data integrity and to facilitate better performance.
* Used Advanced Querying for exchanging messages and communicating between different modules.
* System analysis and design for enhancements Testing Forms, Reports and User Interaction.
* Used Spark Streaming to divide streaming data into batches as an input to spark engine for batch processing.
* Worked on analyzing Hadoop cluster and different Big Data analytic tools including Pig, hive, HBase, Spark and Sqoop.

Environment: Hadoop, HDFS, Hive, MapReduce, Impala, Sqoop, SQL, Informatica, Python, Flume, PySpark, Yarn, Pig, Oozie, Linux, AWS, Tableau, Maven, Jenkins, Cloudera, SAS (BI & DI),PL/SQL, Autosys, Oracle, Sql Server,No Sql,