

**Angad**

**Big Data Engineer**

**Email: angad.prabhav1312@gmail.com Phone: 502-461-0774**

**PROFESSIONAL SUMMARY:**

* 7+ years of professional experience in Statistical modeling, Machine Learning, Data Visualization and Big data.
* Expertise in transforming business requirements into building models, designing algorithms, developing data mining and reporting solutions that scales across massive volume of unstructured data and structured.
* Excellent understanding **of Hadoop Architecture** and Daemons such as HDFS, Name Node, Data Node.
* Data Node, Job Tracker, Task Tracker and Map Reduce Concepts. Hands on experience in installing, configuring and using Hadoop ecosystem components like **Hadoop, HDFS, MapReduce Programming, Hive, Pig, Sqoop, HBase, Impala, Oozie, Zoo Keeper, Spark, SOLR with Cloudera distribution**.
* Hands on experience in various big data application phases like data ingestion, data analytics and data visualization.
* In-depth understanding of Spark Architecture including Spark Core, Spark SQL, Spark Streaming.
* Design and implement database solutions in **Azure SQL Data Warehouse, Azure SQL**
* Implement ad-hoc analysis solutions using **Azure Data Lake Analytics/Store, HDInsight**
* Extensive experience in **Natural Language Processing (NLP)** like Sentiment Analysis, Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating Data Visualizations using R, Python and Tableau.
* Experience in developing different Statistical Machine Learning, Text Analytics, Data Mining solutions to various business problems and data visualizations using R and Python.
* Experience in integrating data, profiling, validating and data cleansing transformation and data visualization using R, SAS and Python.
* Extensive working experience with Python Libraries including **Scikit-learn, Pandas, NumPy, H20 and Pyspark**.
* Expert in data ingestion tools like **Sqoop, Flume, Kafka, Spark Streaming**.
* Experience in**Big Data** cleansing scripts like Spark, MapReduce and Pig and developing customized UDF’s in java to extend **Hive** and **Pig Latin** functionality.
* Intensive experience in Hive and exposure on NoSQL Db’s like **HBase, Cassandra** and Mongo DB.
* Ingested data from different sources like Oracle, Teradata, SQL server.
* **Data Migration** and **Data generation** in big data ecosystem.
* Experience in managing and reviewing **Hadoop log files**
* Experience in developing pipelines in spark using **Scala** and **python**.
* Developing streaming pipelines using Kafka and Storm.
* Orchestrated multiple Hadoop application jobs using Sqoop and implementing optimization techniques in Hive, Spark.
* Experience in Python and shell scripting
* Experience working with cloud tools like **Amazon Web Services** and Azure.
* Experience in creating **CI/CD pipelines** using **Jenkins.**
* Worked on **AWS** Lambda, AWS S3 and AWS EMR
* Experienced with the Apache Spark improving the performance and optimization of the existing algorithms in Hadoop using Apache Spark Context, Apache Spark-SQL, Data Frame, Pair RDD's, Apache Spark **YARN.**
* Worked with Apache Spark which provides fast and general engine for large data processing integrated with functional programming language Scala.
* Hands on experience in Sequence files, combiners, Counters, Dynamic Partitions, bucket for the best practice and performance environment.
* Used Scala and Python to convert Hive/SQL queries into RDD transformations in Apache Spark.

## TOOLS AND TECHNOLOGIES:

|  |  |
| --- | --- |
| **Programming Languages:** | **Python:** CARET, glmnet, forecast, XG boost, Sci-kit learn  **SAS:** Forecast server, SAS Procedures and Data modeling  **Spark:** MLlib  **SQL:** Analytical & Windowing functions, Subqueries, joins, DDL/DML statements  **R**: CARET, Random Forest |
| **Big Data Ecosystem** | Hadoop 1.x/2.x(Yarn), HDFS, MapReduce, Pig, Hive, HBase, Tez, Sqoop, Impala, Oozie, Zookeeper, Apache Spark, Spark SQL, Spark Streaming, Apache Kafka, Apache Flume, Cassandra. |
| **Hadoop Distribution** | Horton Works, Cloudera, AWS S3, Lambda, AWS EMR |
| **Databases** | MS-SQL, Oracle, NoSQL, MySQL |
| **Statistical Methods** | **Supervised Learning:** Linear/Logistic Regression, Lasso, Ridge, Decision Trees, Ensemble Methods, Random Forests, Support Vector Machines, Gradient Boosting, XGB, Neural Networks.  **Unsupervised Learning:** Principal Component Analysis (PCA), K-Means, Hierarchical Clustering, Market Basket Analysis, Collaborative Filtering and Low Rank Matrix Factorization  **Sampling Methods:** Bootstrap sampling methods and Stratified sampling  **Model Tuning/Selection:** Cross Validation, AIC/BIC Criterions, Grid Search and Regularization, Dimension Reduction |
| **ETL/BI Tools** | Tableau, Talend, Microsoft SQL Server, Advanced-Excel, ggplot2 |
| **Natural Language Processing/**  **Text Mining** | Document Term Matrix (DTM), Stemming, Lemmatization, Word Embedding,  Semantic, Term Frequency, Dependency Parsing, Sentiment, Natural Language  Generator (NLG), Word Cloud, Named Entity Recognition (NER), Parts Of  Speech(POS) Tagging |
| **IDE** | Spyder, RStudio, Jupyter, Anaconda, H2O, Pyspark, Flask, Django, Docker, R shiny |
| **Other:** | GIT, Statistics, Microsoft Azure, Google Cloud Platform, Amazon AWS, Hadoop |
| **Operating Systems** | Windows, LINUX, Macintosh HD, CentOs, RedHat |
| **Methodologies** | Agile/Scrum, Waterfall |

**CERTIFICATIONS:**

* **AWS Certified Developer Associate**
* **AWS Certified Data Analytics Specialty**

**PROFESSIONAL EXPERIENCE:**

**Adobe, CA Jan 2021 – Till Date**

**Data Engineer**

**Roles & Responsibilities:**

* Working in agile, successfully completed stories related to **ingestion**, **transformation** and **publication** of data on time.
* Design, document, build, test and deploy data pipelines that assemble large complex datasets from various sources and integrate in a unified way.
* Perform validations and consolidations for the imported data, **Data Migration** and **Data Generation.**
* Ingested data from multiple Databases, servers and storing the data into **HDFS** using **spark-flows**
* Expert in implementing advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache **Spark** written in **Scala**.
* Performing **ETL** operations on the historical data tables and exporting into RDBMS data source (**MSSQL**) using spark-flow with **HIVE**.
* Written **T-SQL** scripts and stored procedure according to the business logic and generated fact and dimension tables.
* Automated and orchestrated the process for getting, storing, transforming the data using CA Automation workload Tool and **OOZIE**.
* Involved in importing the data using elastic search to Hadoop using **Kafka**
* Implemented the **Oozie** job for daily imports
* Provide custom end points to the Front-End team for data source.
* Work closely with team of Data Science to take existing or new models and convert them into scalable analytical solutions.
* Developed **Spark scripts** by using Scala shell commands, Java as per the requirement.
* Involved in teams to analyze the Anomaly detection and ratings of the data using ETL tool **Talend**.
* Involved in converting **Hive/SQL** queries into Spark transformations using Spark RDDs, Scala.
* Wrote Pig Latin Scripts and Hive Queries using Orc schemas to transform the Data sets in HDFS.
* As part of support, responsible for troubleshooting of **Map Reduce Jobs**, Java Scripts, Pig Jobs, Hive
* Worked on performance tuning of Hive & Pig Jobs.
* Used **JDBC** connections to connect the spark applications and MySQL.
* Implemented **Kerberos** security in all environments.
* Develop Spark jobs using **Pyspark** for the whole data pipeline.
* Managed and reviewed Hadoop log files to identify issues when job fails and used HUE for UI based pig script execution, Automic scheduling.
* Involved in creating data-lake by extracting customer's data from various data sources to HDFS which include data from Excel, databases, and log data from servers.
* Designed number of partitions and replication factor for Kafka topics based on business requirements and worked on migrating **Map Reduce** programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
* Used **GIT** as the version control system and worked in UNIX environment
* Worked in **Agile** Methodology.

**Technologies Used:** Hadoop, Cloudera, **Map Reduce**, Spark, Hive, Apache NiFi, Pig, Sqoop, Shell Scripting, Storm, Talend, Kafka, Oracle, Teradata, AWS Lambda, EMR, MySQL, Python, JIRA.

**JP Morgan Chase Bank, NJ May 2021 – Dec 2021**

**Big Data Engineer**

**Roles & Responsibilities:**

* Responsible for building scalable distributed data solutions using Hadoop.
* Worked on Apache Spark using Python.
* Extensively used Spark-core and spark-SQL libraries to preform transformations on the data.
* Design and develop ETL integration patterns using Python on Spark.
* Develop framework for converting existing mappings to PySpark(Python and Spark) Jobs.
* Create Pyspark frame to bring data from RDBMS Database to Hadoop.
* Translate business requirements into maintainable software components and understand impact (Technical and Business)
* Provide guidance to development team working on PySpark as ETL platform
* Makes sure that quality standards are defined and met.
* Optimize the Pyspark jobs to run on Production Cluster for faster data processing
* Provide workload estimates to client
* Implement CICD(Continuous Integration and Continuous Development) pipeline for Code Deployment
* Reviews components developed by the team members
* Used different transformations and functions.
* Worked on performance tuning on the existing spark applications.
* Continuous monitoring and managing the Hadoop cluster through Yarn UI.
* Experienced in writing shell scripts to process the jobs.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark.
* Scheduled Spark jobs by using Autosys tool

**Technologies Used:** Spark, Spark SQL, Oracle, SQL Server, HDFS, YARN, IntelliJ Idea, Maven, Cloudera, Hive, Shell Scripting, Python, Pyspark

**Duke Energy, NC April 2019 – April 2021**

**Big Data Engineer**

**Roles & Responsibilities:**

* Responsible for ingesting data from different sources into **Hadoop** using **Sqoop**.
* Experienced in handling Incremental daily loads and full loads.
* Involved in creating, loading Hive tables, and analyzing data using hive queries.
* Developed **Hive** queries to process the data and generate the data cubes for visualizing.
* Performed transformations using spark to create intermediate tables in hive.
* Using Pyspark for spark transformation.
* Extensively worked on **spark SQL** to perform transformations.
* Worked on bash script to schedule jobs on both Sqoop, hive and spark jobs.
* Worked on creating the **oozie** jobs such as workflows, coordinators, and job properties to reflect made changes to run on the edge node.
* Performed manual Sqoop loading for one time loads for the missing data in the already existing hive tables.
* Validating the hive tables to see the schema and column data types to match with the source data.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
* Involved in scheduling Oozie workflow engine to run multiple Hive jobs, spark jobs.
* Involved unit testing, interface testing, system testing and user acceptance testing of the workflow tool.
* Responsible for building scalable distributed data solutions using Hadoop cluster environment with Hortonworks distribution.
* Used **Amazon S3** as a Data-lake to the data pipeline running on spark.
* Worked on writing **AWS Lambda** function which creates the **EMR** cluster and auto terminates thelambda after the job is done.
* Adding the steps to the **EMR** cluster within the bootstrap actions of **AWS Lambda**.
* Worked on fetching data from various source systems such as **HIVE** and **Amazon S3.**
* Used **GIT** as the version control system and worked in UNIX environment
* Worked in **Agile** Methodology.

**Technologies Used:** Spark, Spark SQL, Scala, Python, Hive, Sqoop, oozie, bash scripting, Oracle, SQL Server, Tera data, HDFS, YARN, Jenkins, IntelliJ Idea, Maven

**Shutterfly Inc, AZ FEB 2018 – March 2019Data Engineer/ Scientist**

* Performed Data Profiling to learn about behavior with various features such as traffic pattern, location, and time, Date and Time etc. Integrating with external data sources and APIs to discover interesting trends.
* Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.
* Experience in installation, configuring, supporting and managing Hadoop Clusters using Apache, Cloudera (CDH 5.X) distributions.
* Worked on Cloudera distribution for Hadoop ecosystem and installed and configured Flume, Hive, Pig, Sqoop and Oozie, Automic on the Hadoop cluster.
* Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark **YARN**.
* Built **Machine Learning** models to identify fraudulent applications for loan pre-approvals and to identify fraudulent credit card transactions using the history of customer transactions with supervised learning methods.
* Develop Oozie workflows to schedule the Scripts on a daily basis.
* Develop Spark jobs using **PySpark** to create a generic framework to process all kinds of flat files.
* Performed Data Cleaning, features scaling, featurization, features engineering.
* Managed and reviewed Hadoop log files to identify issues when job fails and used HUE for UI based pig script execution, Automic scheduling.
* Involved in creating data-lake by extracting customer's data from various data sources to HDFS which include data from Excel, databases, and log data from servers.
* Used various Spark Transformations and Actions for cleansing the input data and involved in using the Spark application master to monitor the Spark jobs and capture the logs for the spark jobs.
* Experience in refactoring the existing spark batch process for different logs written in Scala.
* Implemented**Big Data** tools like Spark using Scala and utilizing Data frames and **Spark SQL** API for faster processing of data and worked on extensible framework for building high performance batch and interactive data processing application on hive.
* Extracted Real time feed using Spark streaming and convert it to RDD and process data into Data Frame and load the data into Cassandra.
* Worked on fetching data from various source systems such as **Amazon S3.**
* Customer segmentation based on their behavior or specific characteristics like age, region, income, geographical location and applying Clustering algorithms to group the customers based on their similar behavior patterns.
* The results from the segmentation helps to learn the Customer Lifetime Value of every segment and discover high value and low value segments and to improve the customer service to retain the customers.
* Performed **Clustering** with historical, demographic and behavioral data as features to implement the personalized marketing that offers right product to right person at the right time on the right device.
* Addressed over fitting and underfitting by tuning the hyper parameter of the algorithm and by using **L1 and L2 Regularization**.

**Technologies used:** Hadoop stack, Spark SQL, KSQL, Spark-Streaming, Scala, Cassandra, Cloudera, Kafka, Hive, Pig, Sqoop, Linux.

**GIIR America Inc, NJ OCT 2016 – JAN 2018**

**Software Engineer**

* Extensively involved in all phases of data acquisition, data collection, data cleaning, model development, model validation, and visualization to deliver data science solutions.
* Extracted data from database, copied into **Hadoop Distributed File system (HDFS)** and used Hadoop tools such as Hive and Pig Latin to retrieve the data required for building models.
* Developed, Installed and configured Hive, Hadoop, Bigdata, hue, Oozie, pig, Sqoop, Kafka, Elastic Search, Java, J2EE, HDFS, XML, PHP and Zookeeper on the Hadoop cluster.
* Worked on data cleaning and ensured data quality, consistency, integrity using Pandas, NumPy.
* Tackled highly imbalanced Fraud dataset using sampling techniques like down-sampling, up-sampling and SMOTE (Synthetic Minority Over-Sampling Technique) using Python **Scikit-learn**.
* Worked on loading AVRO/PARQUET/TXT files in Spark Framework using Scala language and created Spark Data frames and RDDs to process the data and save file in parquet format in HDFS to load into fact table using ORC Reader.
* Migrated Map Reduce programs into Spark transformations using Scala.
* Implemented a Python-based distributed random forest via **PySpark** and **MLlib**.
* Used **AWS S3**, **DynamoDB, AWS lambda, AWS EC2** for data storage and models' deployment.
* Created and maintained reports to display the status and performance of deployed model and algorithm with Tableau
* Implemented CICD allowing for deploy to multiple client Kubernetes/AWS environments.
* Worked on Hive to implement Web Interfacing and stored the data in Hive external tables.
* Implemented Hive Partitioning and Bucketing on the collected data in HDFS.
* Involved in Data Querying and Summarization using Hive and created UDF’s, UDAF’s and UDTF’s.

**Technologies used:**Hadoop stack, Spark SQL, KSQL, Spark-Streaming, Java, J2EE, **AWS S3**, **AWS EMR**, google cloud, GraphX, Scala, Python, Pyspark, Kafka, Hive, Pig, Sqoop, Oozie, vertica, Impala, CICD, Cassandra, Cloudera, Oracle 10g, MySQL, spring boot, Linux.

**Blue Cross Blue Shields, TX SEPT 2015 – OCT 2016   
 Jr. Software Engineer/Intern**

* Performed advanced and predictive data analytics using data science technology to predict either medical claim is legit or fraud by using very effective and power machine learning algorithms.
* Worked on a POC for extracting real time data using Kafka and spark streaming by Creating
* DStreams and converting them into RDD, processing it and stored it into Cassandra.
* Imported data from RDBMS systems like MySQL into HDFS using Sqoop.
* Worked on different file formats (ORCFILE, Parquet, Avro) and different Compression Codecs (GZIP, SNAPPY, LZO).
* Experience with CDH distribution and Cloudera Manager to manage and monitor Hadoopclusters.
* Developed Oozie Bundles to schedule pig, Sqoop and hive jobs to create data pipelines.
* Created Data Pipelines as per the business requirements and scheduled it using OozieCoordinators.
* Used Spark-SQL to Load JSON data and create SchemaRDD and loaded it into Hive Tables and handled Structured data using Spark SQL.
* Used Spark-SQL to Load data into Hive tables and Written queries to fetch data from thesetables.
* Processed the Raw data from CSV files in to organized form by applying Data Cleaning techniques using **Pandas** and **NumPy**.
* Implementing MVC Design pattern for the Application.
* Performed analysis using **Count Vectorizer**, **TF-IDF,** Linear SVC and pipelined those to develop model to predict spam and classify which emails to respond back and which to put in the junk emails.
* Built **Confusion Matrix** and **Classification** report to see the performance of the algorithm.
* The jobs were made to run successfully by solving data quality issues using SQL, efficient coding practices, macros and stored procedures.

## EDUCATION:

Master of Science

Trine University

Business Analytics

Bachelor of Science

California State University, Long beach

Applied Mathematics with Economics Concentration