CHAYA LAKSHMI PARUCHURI

Working as Software/ML/Data Engineer in JP Morgan Chase.

chayalakshmiparuchuri270@gmail.com, Phone: +91-8328325969

Career Objective

To obtain a creative and challenging position that gives me an opportunity for self-improvement, while contributing to the symbolic growth of the organization with my technical, innovative and logical skills.

Skills and Experience Summary

- Working experience on **Hadoop technologies** like Map/Reduce, Hive, Pig, Hbase Sqoop, Oozie, PySpark.
- **AI/ML model implementation** using **pySaprk** includes **data extraction, data ingestion**, **data pre-processing and model execution** followed by model score distribution.
- Working experience on programming languages python.
- Strong technical knowledge in **Hadoop** and **spark** solution development.
- Possess good interpersonal skills, goal oriented and always interested to learn new things and critical technologies.
- Acquired the top competencies in Hadoop and python in TCS.
- Worked in Telecom and **Health Care domains**.
- Excellent communication skills, problem solving skills, an attitude to learn the new cutting edge technologies.

Technical Skills

Big data	Apache **Hadoop**, **Hive**, **Spark**, PIG, Map Reduce, Sqoop, Oozie, Hbase
Languages	**Python**, Java(J2SE), **pySpark**
Operating system	UNIX, Linux, Windows
Database	SQL Server and SQL, PL/SQL

Certifications in coursera

Big data and spark	Big Data Essentials:HDFS, Mapreduce and SparkRDD, Big Data Analysis:Hive,Spark SQL,DataFrames and GraphFrames.
Python	Applied Data Science with Python:Specialization

Project Details of Current Employer(JPMC): From Sep 2020 to till date

Domain	**Banking**
Technology	Pyspark
Project Type	AI/ML Migration project
Description	**Implementation of ML model using pySpark in distributed system, Hadoop framework.**
Role/Responsibility	• Involved in end to end implementation of the ML model projects ,working on model implementation, **writing unit test cases** and prod deployment using **CICD** tools (jenkins/jules) . • Writing pySpark scripts to make them utilise the distributed nature of spark. • Worked on **model implementation** of Credit Cards Risk analysis and auto loan models, contributed in end to end implementation of data engineering pipeline and model score distribution. ➢ **Extraction** of historical data (TB's) from multiple sources. ➢ **Ingestion** of data into hive tables using inhouse frame work. ➢ Creation of **custom** variables using complex transformations and amalgamation of data to create central dataset. ➢ Segmentation of data by using **pd score** equations (**logistic regression**) and **decision trees**. ➢ Calculation of standard and advanced rwa and insertion of data into partitioned hive tables. • Orchestration of entire process using inhouse orchestration framework with help of configuration file. • Performance tuning of the model by repartitioning the data at the module level ,using parquet formats for saving intermediate outputs and increasing memory and executor configurations in spark-submit.

Domain	**Banking**
Technology	Pyspark
Project Type	Module Development (ML Model enhancements)
Description	The objective is to implement a generic module to mark specific loans based on the reference data
Role/Responsibility	• Involved in end to end implementation of model enhancements, worked on change development, unit testing, integration testing and prod deployment. • Implementation of two functionalities to handle the requests at both input level and output level. • Output level functionality involves checking various variable and complex multiple conditions and changing the output variables based on the results. • Input level functionality involves changing the input variables with the respective values based on the lookup table. • This module can be plugged into any model and can be used. • **Optimised** the module in such a way that execution of this module for any kind of dataset will not take more than 2 mins.

Project Details of Previous Employer(Optum): From July 2017 to Sep 2020

Domain	**Healthcare**
Technology	pyspark, pytesseract
Project Type	Engine Development
Summary	The objective is to extract the text from clinical charts in pdf format and ingest the data into hbase.
Role/Responsibility	• Played role of Developer. • Used python packages to identify if the pdf is electronic or scanned . • Developed functions to extract text from electronic pdf directly and used pytesseract ocr to get text from scanned images. • Implemented module to handle images of all sizes without throwing memory error in spark. • Improved the performance of the developed engine to process GB's of data in few minutes. • Ingestion of text data into **hbase** using **spark**. • Involved in cycles of development and testing.

Domain	**Healthcare**
Technology	**pyspark**,**pandas**
Project Type	Module
Summary	The objective is to create user specified no of parquet files for a given batch of text files and generate the stats for the same.
Role/Responsibility	• Takes the list of batches and no of parquets to be generated as inputs. • Distributes the files in the batches equally among all executors based on file count in each folder. • Speedup the creation of parquet files for huge data by utilising spark parallelism. • Generation of page stats table with processed,missed,errored and total page counts. • Involved in cycles of development and testing.

Domain	**Healthcare**
Project Name	**Chart value score**
Technology	pyspark,oozie,python,hbase
Project Type	Application Development
Summary	The objective of this module is to generate the chart value score for chart which can be used to prioritize for manual coding of chart.
Role/Responsibility	• Played role of Developer in this project. • Developed python and **pyspark** scripts to perform following operations: ➢ Extract required data from hive tables using **spark** as first input data set and acquire data from files as second data set. ➢ Cleansing and deduplication of input datasets. ➢ Applying transformations like hierarchy filtering and categorise the hcc's identified across both datasets ➢ Calculation of raf score for the required categories and grouping of the charts into 7 groups based on range of raf score. • Automated the entire process using oozie workflow.

Domain	**Healthcare**
Project Name	**Chart Abstraction**
Technology	**Python** opencv,tesseract-ocr,pandas
Project Type	POC
Summary	The objective of this **POC** is to generate a table from the content of chart which is in form of image.The chart contains various entities like textboxes,checkbox and tables.
Role/Responsibility	• Played role of Developer in this project. • Developed python module using opencv package and template matching concept to crop the image into multiple blocks. • Pass each block through **OCR** and write the content to file. • Detection of checkbox based on the pixel intensity inside checkbox. • Extraction of data inside tables based on character and number recognition.

Project Details of Previous Employer(TCS): From July 2014 to July 2017

Client	**Healthcare client**
Project Name	**Bigdata project**
Technology	**Hive**,Sqoop,Oozie,Netezza,SAS,**python**,**hbase**
Project Type	Migration project
Summary	The main goal of this project is migration from SAS environment to Hadoop for minimising the execution time of prediction models running in SAS.
Role/Responsibility	• Played role of Developer in this project. • Importing the data from netezza to hive using sqoop. • Understanding the Business logic of macros in SAS. • Develop the hive scripts implementing the same business logic as that of source. • Implementing hive performance tuning techniques to reduce the execution time. • Developing UDF's for implementation of critical functionalities. • Unit testing and Integration testing of the code by matching the counts of records between source and target systems. • Creation of Oozie workflow as part of automation process. • Used python to develop the automation for importing the tables,triggering the oozie workflow and logging the errors and status of the process. • Logging information was stored in hbase tables.

Client	**Healthcare client**
Project Name	**Bigdata POC**
Technology	Mapreduce,Hbase,Netezza.
Summary	Usually the Hbase data is loaded into hive tables using **Hbase-hive** integration and exported to netezza.But when the data has multiple versions for the same unique key,only first record is getting inserted into hive table.Also if two column families have same column,the data of second column is being missed in hive tables.This POC aims at exporting the data to netezza directly from hbase including multiple versions of records and all columns from all column families.
Role/Responsibility	• Understanding the input data. • Develop MR program to read the hbase tables and implemented logic by creating the hash maps with column family as key and version,column family,column name and column value as value. • Write each column family data into each hdfs file. • Export the hdfs files to netezza.

Client	**Cablevision**
Project Name	Cablevision
Technology	**Pig, Sqoop**, Shell scripting, Amazon Redshift
Project Type	Migration project
Summary	To migrate the database from oracle and netezza to Amazon Redshfit by implementing the existing procedures and other features of oracle and netezza in Redshift.
Role/Responsibility	• Played role of Developer in this project • Understanding the Business logic of procedures in oracle and netezza. • Develop the code for procedures of existing databases in Redshift implementing the same business logic as that of source. • Develop various Pig & Hive scripts for moving the informatica code to redshift by performing the data transformations in Pig • Unit testing and Integration testing of the code by matching the counts of records between source and target systems.

Client	**Comcast Cable**
Project Name	Comcast Cable Communications Management(POC)
Technology	Hadoop, Pig, Sqoop, Shell scripting, Java
Project Type	Application Development
Summary	To find out the Victory Tool capability to Integrate with Hadoop Technology and leverage the Benefit for carrying out the ETL Process that includes Transformations, N-Way Reconciliations and perform Lookups
Role/Responsibility	• Played role of Developer in this projects • Understanding the Business & Functional specifications use case. • Develop the system to extract data from the source systems into the Hadoop environment. • Develop various Pig scripts for performing the data transformations and Performing Reconciliations logics. • Unit testing and Integration testing of the system in Hadoop Cluster.

Qualifications:

(In chronological order starting from the most recent)

Degree and Date	Institute	Percentage	Major and Specialization
B.TECH - 2014	VR Siddhartha Engineering college,Vijayawada	8.82	Computer Science and engineering
+2 - 2010	Sri Chaitanya junior college,Vijayawada	97.5	MPC
10th - 2008	Vijayasri Sunflower Public school,Challapalli	92.16	

Personal Details:

Date of Birth	12-10-1992
Nationality	Indian

Declaration

I hereby declare that the particulars furnished above are true to the best of my knowledge and belief.

Place: Hyderabad Signature

Date: 06 June 2022 (Chaya Lakshmi Paruchuri)

`