

## PALLABI CHAKRABORTY

469-664-3787\* [pallavichakraborty@yahoo.com](mailto:pallavichakraborty@yahoo.com)

GitHub: <https://github.com/pallabichakraborty>

Linkedin: [www.linkedin.com/in/pallabi-chakraborty](http://www.linkedin.com/in/pallabi-chakraborty)

### SUMMARY PROFILE

Master's in computer science from US University along with **8+ years of experience working on Hadoop Ecosystem and ETL Processing** using Oracle 10g and IBM DB2 and DataStage. Core Technical skills include SQL, PLSQL and Big Data with experience in languages like Java, Python and R. Hands on experience in all stages of SDLC. Experienced in managing technical teams. Experience with Multi cloud (Oracle and Google) and familiar with Hybrid cloud infrastructure. Areas of interest include Data Science, Artificial intelligence along with processing and transformation and analysis of data.

### TECHNICAL SKILLS

Big Data Ecosystems	Hadoop, MapReduce, HDFS, HBase, Hive, Pig, Sqoop, Cassandra, MongoDB
Databases	NoSQL, Oracle 9i, 10g, MySQL, DB2 10.1
Programming Languages	Python, Oracle SQL, Oracle PL/SQL, IBM DB2
ETL Tools	IBM DataStage
Reporting & Visualization	Tableau, ReportServer
Oracle Cloud	Oracle Cloud Services like (Autonomous Database, Oracle Machine Learning)
Google Cloud	Tensorflow, Data Lab, Big table, Big Query

### PROFESSIONAL EXPERIENCE

#### **insITe**

**Nov 2019 – June 2020**

#### *Backend and Reporting Intern*

- Architected and created new databases for migration which included implementation of new data loads and a migration from a denormalized database to a normalized reporting database structure, improving efficiency by 25%. Created algorithm to synchronize the production database to the reporting server hourly.
- Created reconciliation scripts to facilitate troubleshooting of the migration issues, reducing manual testing.
- Setup development and production instances for ReportServer and created reports which can be used by the decision makers using the entered data by the operations teams.

#### **Rooster Consulting**

**May 2017 – Jan 2018**

#### *Intern*

- Created front-end webpage for placement of consultants and integrated with the back-end utilizing HTML, JSReact and Javascript.

#### **NYU Tandon School of Engineering**

**Sep 2016 – Dec 2016**

#### *Graduate Assistant – Computer Security*

- Tutored students with basic concepts of computer security and assisted the professor with grading and planning homework.
- Researched computer security topics to create presentations and activities for 100 undergraduate students.

#### **Infosys Ltd., Bangalore, India**

**Dec 2011 – July 2015**

#### *Technology Lead, Technology Analyst*

- Led a team of 6 developers to implement **Big Data** System used to distribute compensations to the farmers by Government of India. The system authenticates beneficiaries and field users using biometrics.
- The system in production uses Apache HBase database hosted on **32 node Hadoop Cluster** as the data store. This resulted in significant increase in security for the system prone to identity theft.
- For Proof of Concept, installed 4 node Hadoop Ecosystem with Apache HBase, Pig and Phoenix on a test server. Apache Hive configured using MySQL database.

- Evaluated **Hortonworks, Cloudera** as well offerings from IBM and Teradata as an exercise to choose the distribution for production deployment.
- Implemented **MapReduce** Jobs for highly data intensive jobs to save on the conversion time when using higher complexity Hadoop Ecosystem members.
- Led a team of 7 developers to create automated SLA system provides near to real time SLA information and alerts in case of SLA breaches to be used by the operations team using IBM DB2, reducing SLA reporting from days to an hour.
- Led a team of 4 developers to build an in-house accounting system which would retrieve the taxation data for the tax payments of employed citizens of India, 25 million citizens. This implementation performs the accounting of the direct taxes paid by the Indian employed population using **IBM DB2**.
- Worked on high volume data migration, which ranged up to 3 billion records for the bigger sources. The migration activity done to migrate data from an **Oracle 10g** database to IBM DB2 database.
- Worked on development of the Processing Module, which traverses through at least 500 million records at a given point of time and generates data to be used for the tax filing by the Indian taxpayers.
- Awarded Technology Lead of the Month in Infosys for effectively managing the deliverables along with grooming the team to be able to work effectively as well as grow to the next levels.

## **Quintiles Technologies India Pvt. Ltd., Bangalore, India**

**July 2007 – Dec 2011**

*Database Programmer/Analyst, PL/SQL Programmer, Associate Programmer*

- Worked on Automation of the customer survey module, process compliance module, project review dashboard for the projects being handled by Operations team, using Oracle Application Express; which was earlier being done using excel sheets. This resulted in improving the efficiency by almost 75%.
- Worked on **Cognos** Reporting Studio to generate data tracking reports using patient data. This data is used by clinical research associates to verify the efficacy of the drug under clinical trials.
- Worked on generating **SAS reports** and SAS datasets for the usage of the SAS statisticians to generate reports for the safety of the medicines being used in clinical trials.
- Worked on Data validation procedures to determine the data consistency of the patient data being collected in clinical trials, which are submitted, to FDA for approval of the drugs.
- Created front-end screens to be used in the front-end screens of clinical trials using **Oracle's Inform Architect** and **Central Designer**. Created and tested front-end data validation checks.
- Awarded Star Programmer of Year in Quintiles for delivering deliverables with minimal defect ratio.

## **EDUCATION**

NYU Tandon School of Engineering, Brooklyn, NY  
Master of Science, Computer Science

Dec 2016  
**GPA: 3.75**

Visvesvaraya Technological University, Bangalore, India  
Bachelor of Engineering, Electronics and Communication

Mar 2007  
**GPA: 3.5**

## **CERTIFICATIONS**

1Z0-047	Oracle Database SQL Expert
1Z0-147	Oracle Database 10g: Program with PL/SQL
DB2 10.1 Fundamentals	IBM Certified Database Associate - DB2 10.1 Fundamentals
1Z0-931-F	Oracle Autonomous Database Cloud 2019

## **ACADEMIC PROJECTS**

### **Analysis of Sample Sales Data to deduce patterns**

Employ visualization of sample sales data to generate patterns to deduce if giving discounts results in losses and provide recommendations on how to approach to get better profits

Implementation Details:

- Tableau story points developed consisting of dashboards which provide a better insight to the data.

Tools & Technologies: Tableau Desktop

### **Data Analysis - Yelp 2016 challenge dataset**

This project involved analysis of Yelp data using Yelp 2016 challenge dataset: [https://www.yelp.com/dataset\\_challenge](https://www.yelp.com/dataset_challenge).

This project involved emerging with trends and conclusions using the data available.

Implementation Details:

- R modules developed to analyze the Yelp Data to deduce useful inferences.
- Visual representation of the data for better understanding and analysis created using additional R packages like ggplot2

Tools & Technologies: Apache Spark, R, MongoDB

### **Web Crawler**

This involved creation of a WebCrawler which can crawl through the internet using the first 10 links from the Google Search on a particular search term or group of terms. Two variants of web crawler were created - Focused Crawler as well BFS Crawler

Implementation Details:

- The input provided to this module is one or multiple words in command line interface.
- The module hits the google search engine utilizing the Google API Python module.
- Using the first 10 links fetched, the module fetches the links related to the search terms in the corresponding pages and then recursively performs the same task on the links obtained during the page content fetch.
- To prevent an infinite fetch of the same pages done by various developers to undermine the search engines, the utility only checks till 10 levels of pages.

Tools & Technologies – Python

### **Analysis of Google Flu Trends Data**

This involved understanding and analysis of Google Flu Trends and usage of the same to arrive to conclusions.

Implementation Details:

- Data collected by Google, from various countries to compare flu trends
- The data was utilized to generate predictions for future time using **R packages**

Tools & Technologies: R, R Studio

### **Can social unrest be predicted using Social Media**

This involved analysis of social media especially Twitter feed to deduce whether social unrest can be predicted by exploiting the tweets and respective emotions involved.

Implementation Details:

- This project evaluates if we can predict an event by analyzing social media such as tweets. Python module was developed to download the twitter feeds due to the cap on the download of tweets by an individual account.
- The feeds which were downloaded were fed to Machine learning packages provided in R to forecast future events.
- The module generated forecasts with a precision rate of around 50% at the point of submission, this can be further enhanced by adding further data points.

Tools & Technologies – R, R Studio, Python, Twitter Feeds

### **Predicting spam websites in Web Search Engines**

This project involved usage of *machine learning* in order to classify if a given website is a spam or non-spam using its link, structure, meta and content. The training data used consist of already classified website links.

Implementation Details:

- Java modules were developed which utilizes already classified spam website data to classify any given website link as spam or non-spam.
- The classification is done on the basis of content of websites and structure of the website links. This is done using Random Forest and Decision Tree Machine Learning Techniques provided by Pyspark.

Tools & Technologies – Java, Python, Apache Spark