# Professional Profile:

# Data engineer with 8+ years of experience with expertise in full life cycle software Development, Design, and Implementation of Data Warehousing, and Business Intelligence applications.

* **Strong programming** background with proficiency in various languages like **PYTHON, Unix Shell scripting, SCALA etc.**
* Designed, Developed and Implemented complex ETL jobs **using Informatica and DB2** for Data cleansing, Data Transformation, Data Integration, Data Profiling and involved in Performance Tuning of jobs
* Strong experience in **Big Data** solutions using various tools in **Hadoop ecosystem** like Hive, Sqoop, Spark (RDD, Spark SQL), Kafka, Flume.
* Expertise in **data movement** from various sources to HDFS and data transformation using Hadoop tools like **HIVE,HDFS, SQOOP** and data processing toolslike **APACHE SPARK(1.6/2.x)**.
* Extensive experience on **RDD and Dataframe** architecture and utilizing the concepts to build/optimize various spark applications,
* **End to End** Expertise in **data warehouse domain** with knowledge of data modeling and using ETL tools like **Informatica 9.0.**
* Strong knowledge of **ETL methods** for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis using ETL tools like **Informatica** **Power Center 10.x/9.x/8.x**, **UNIX shell scripting** and databases like **DB2, Postgres, Oracle, MS SQL Server and Teradata.**
* Experience in using **job monitoring and scheduling** tools like oozie, Control-M, Crontab.
* Excellent in analyzing data using **SQL, HiveQL**, **UNIX shell scripting**.
* Understanding on **Safe Agile concepts** and worked using Agile Methodology for SDLC (Software Development Life Cycle) and utilize **agile scrum** meetings for identifying stories and creatively working on them.
* Experience in using **GIT Lab** for code version control and using **Version One** for agile lifecycle management.
* Researching on A**mazon Web Services (AWS) cloud services** like EC2, S3, etc. for hosting existing data in future.
* Exploring on **Machine learning** concepts.
* Strong **Impact analysis** and r**equirement gathering** experience acquired by constant communication with Client, Technical teams and functional Business Analysts.
* Worked in **offshore-Onsite model** for various project releases with onsite experience in project implementation and requirement gathering.
* **Excellent** troubleshooting, problem solving, team working skills.
* High level of integrity and dedication to quality, excellence and corporate objectives.
* **Self-motivated** with the ability to work in fast paced environments and a believer in delivering results.
* **Lead & mentored team** for various project deliveries ensuring high quality in development, testing and implementation of various releases in onsite-offshore operating mode.
* Received trainings in **“Cloudera data analyst training” and “Developer training for Apache Spark and Hadoop”** courses from Cloudera.

**Education:**

* Bachelor of Engineering (B.E) in Electronics Engineering from Nagpur University.

# Technical Skills:

**Big Data tools:**  Hive, HDFS, Sqoop, Impala, Beeline, HUE, Spark 1.6/2.1

**ETL tools:** Informatica 10.x/9.x/8.x.

**Languages:**  PL/SQL, Korn Shell Scripting, Python 2.7/3.

**Database:** IBM DB2, Postgres 9.5, Oracle 10, Oracle-9i, Teradata V2R5

**OS/Environment:** UNIX, Windows.

**Job scheduler:**  Crontab, Control-M

**Project Methodology:**  SAFE AGILE and SDLC Waterfall methodology.

**Project management:** Version One, GIT

**Data pipeline:** GIT CI, Jenkins

**Service Management:** HP Service Manager

# Experience:

* + - * 1. **Statefarm Mutual Automobile Insurance Company**

##### Project : Claims Machine Learning And Artificial Intelligence

## Role : Data Analyst

## Tools : Hadoop (Hive, Sqoop, Spark), Python, Unix Shell scripting,

## Duration : July 2018 till date

## Location : Bloomington, IL

*Overview****:***

State Farm is a family of insurance and financial services companies that together serve tens of millions of customers in the U.S. Our many lines of business together offer over 100 products. This project aims to analyze existing massive amount of data and help building data models based on requirement for better customer experience in insurance policies and claims. To achieve this, we have used Hadoop tools like Hive, HDFS, Sqoop, scripting languages like Python, HiveQL and databases like DB2, Postgres.

*Responsibilities****:***

* Worked on creating Hive tables and written Hive queries for data analysis of massive data sets to meet business requirements
* Experienced in data extraction from IBM DB2 into HDFS using HiveQL, SQL and shell scripting.
* Used **Jupyter Notebook** for creating Python (2.7/3) scripts for data analysis and transformation using various python libraries like Pandas, Numpy, etc.
* Developed spark SQL scripts for exploratory analysis using Apache Spark Dataframes.
* Worked with RDD and dataframes in apache spark for data analysis and transformation.
* Involved in loading and transforming large sets of Structured, Semi-Structured and Unstructured data and analyzed them by running Hive queries.
* Extensively worked with Jenkins team for continuous integration and for End to End automation for all build and deployments.
* Worked with cross functional consulting teams within the data science and analytics team to design, develop, and execute solutions to derive business insights and solve clients' operational and strategic problems.
* Continuous monitoring and managing the Hadoop/spark cluster using Cloudera Manager
* Developed UNIX shell scripts and automated them using CRON job scheduler.
* Experience working in Agile Methodology and used Version One and GIT lab for maintaining the stories about project.
* Worked with architecture and development teams to understand usage patterns and work load requirements of new projects to ensure the Hadoop platform can effectively meet performance requirements and service levels of application
* Researching and implementing learnings on Amazon Web Services (AWS) utilizing tools like EC2, S3, etc. for hosting existing data in cloud.
* Exposed to machine learning (ML) algorithms.

##### Project : Claims Fraud Detection analysis

## Role : Data Analyst

## Tools : Hadoop (Hive, HDFS, Sqoop, Spark, Oozie), Python, Unix Shell scripting, Informatica, Control-M

## Duration : May 2016 to July 2018

## Location : Bloomington, IL

*Overview****:***

State Farm is a family of insurance and financial services companies that together serve tens of millions of customers in the U.S. Our many lines of business together offer over 100 products. Statefarm Insurance optimizes processes for insurance policies. This project aims to analyze existing claims data and create relevant procedures to be performed for better customer experience in insurance policies and claims. To achieve this, we have used ETL tools like informatica, Hadoop tools like Hive, HDFS and Apache Spark, scripting languages like HiveQL, Python and databases like IBM DB2.

*Responsibilities****:***

* Developed and implemented scalable solutions for ingesting data from various sources and processing the data utilizing Big Data (Hadoop) technologies such as Hive, HDFS, Spark and UNIX Shell scripting.
* Extracted data from different data sources into HDFS from DB2 using SQL, HiveQL, UNIX shell scripting.
* Created internal, external, parquet Hive tables and worked on them for data analysis and transformation.
* Exposed to concept of loading data into Spark RDD and Dataframes to do in memory data computation to generate the output response in form of XML or JSON files.
* Exposed to coding in Spark for data processing and optimization of the existing algorithms using Spark Context, Dataframe, pair RDD's, Spark YARN.
* Worked extensively in performance tuning of Hive and Spark applications for setting right parameters, correct level of parallelism and memory tuning for faster processing with high volume data.
* Developed workflow in Oozie to manage and schedule jobs on Hadoop cluster to trigger daily, weekly and monthly batch cycles.
* Developed Unix Shell Scripts and automated them using CRON job scheduler.
* Experience in using Cloudera manager for continuous monitoring of jobs running in cluster.
* Used GIT lab for code version control.

##### Infosys Ltd.

## Client : John Deere

##### Project : Data Processing Support

## Role : Senior Informatica Developer

## Tools : Informatica, IBM Mainframe, UNIX shell scripting

## Duration : Mar 2016 to Apr 2016

## Location : Moline, IL

Overview:

The project involves building of informatica mappings to enhance quality and develop informatica mappings for adequate data delivery to customers.

*Responsibilities:*

* Provided on-call support during and after code promotion as needed.
* Re-engineered on existing mappings to support new/changing business requirements.
* Used mapping, variables/parameters, and parameter files to support change data capture and workflow execution process.
* Modified several of the existing mappings based on the user requirements and maintained existing mappings, sessions and workflows.
* Used various transformations like Source Qualifier, Expression, Aggregator, Joiner, Filter, Lookup, and Update Strategy for Designing and optimizing the Mapping.
* Tuned the performance of mappings by following Informatica best practices and also applied several methods to get best performance by decreasing the run time of workflows.
* Prepared SQL Queries to validate the data in both source and target databases.
* Involved in Unit testing, Integration testing, UAT by creating test cases, test plans and helping Informatica administrator in deployment of code across Dev, Test and Prod Repositories.
* Created test plans and did unit testing for the Informatica mappings and stored procedures.
* Involved in Unit, Functional, Integration and System testing and preparation review documents for the same.
* Creation of sessions and workflows according to the data load in to different systems.
* Validated the Mappings, Sessions & Workflows, Generated & Loaded the Data into the target database.
* Monitored batches and sessions for weekly and monthly extracts from various data sources to the target.

##### Larsen & Turbo Infotech

## Client : Citibank

##### Project : Finance and Data

## Role : Senior Software Engineer

## Tools : Abinitio, Oracle, MS SQL

## Duration : June 2014 to July 2014

## Location : Bangalore, India

Overview:

The project involves building and maintenance of graphs using Abinitio to enhance quality of existing process and to meet business requirements for adequate data delivery to customers.

Responsibilities:

* Monitored batches and sessions for weekly and monthly extracts from various data sources to the target
* Interacted with business users and analysts and gathered and documented the technical and business metadata.
* Generated ETL packages with different data sources and loaded the data into target tables by performing different kinds of transformations using Informatica.
* Managed the metadata associated with the ETL processes used to populate the Data Warehouse.
* Applied various data transformations like Slowly Changing Dimension (SCD), Aggregate, Sort, etc.
* Created complex stored procedures and functions for integrity checks of transferred data.
* Performed unit testing and SQL performance tuning.
* Helped provide end users with the ability to access and utilize the data from target tables.

##### HCL technologies Ltd.

## Client : Citibank

##### Project : FlexiBus

## Role : Senior Software Engineer

## Tools : Ab initio, Oracle, UNIX shell scripting

## Duration : June 2013 to July 2014

## Location : Bangalore, India

Overview:

The objective of FlexiBUS project is to develop, test and implement a robust generic code which caters to 90+ countries across four regions namely NAM, LATAM, APAC and EMEA. An automated and reliable Extract, Transform and Load (ETL) process to extract data from FLEXCUBE and create handoff to partner systems. The FlexiBUS process gathers data from source system interfaces transforms data, creates files based on respective structures of partner systems. The scope in this project is to create handoff files for partners as well as to load the data to FLEXCUBE stage area received through partner feed files.

*Responsibilities:*

* Helped provide end users with the ability to access and utilize the data from target tables.
* Creating Interface contract document, High level design and Low level design documents based on broadcast received from bank.
* Development and design of code using Ab initio 3.12 with extensive use of PL/SQL for analysis and coding.
* Performance improvement of existing applications and newly built.
* Exhaustive testing of applications using SQL scripts in Oracle.
* Preparing various documents such as Unit Test Specifications, System Test cases, SQL scripts.
* Used Korn shell scripts for data loading and execution of mapping
* Reviewing and exhaustive testing of components before production implementation along with groups involved that involves quality analysis, defect logging and resolution.

##### Tata Consultancy Services Ltd.

## Client : ABN AMRO, Netherland

##### Project : Remote Marketing Intelligence

## Role : Software Engineer

## Tools : Ab initio, Oracle, UNIX shell scripting

## Duration : Sep 2012 to May 2013

## Location : Bangalore, India

Overview:

The Financial Business Intelligence data warehouse system for client was implemented to handle the increasing data volumes generated by the growing active customer base, while improving customer service levels. The project involves maintenance of warehouse by means of integrating data from different feeder systems in the form of flat files tables, etc. which is, then processed using ETL procedures before loading to target tables.

Our team was responsible for updating the mappings with new table/column changes, troubleshooting data and load issues of external flat files, performance tunings and performing regular maintenance tasks.

*Responsibilities:*

* Reviewing and exhaustive testing of components before production implementation along with groups involved that involves quality analysis, defect logging and resolution.
* Involved in the requirement definition and analysis in support of data warehousing efforts.
* Involved in designing the Data warehouse using Informatica 8.6 ETL tool by using Source Analyzer, Warehouse Designer, Mapping Designer and Transformations.
* Developed ETL procedures to ensure compliance with standards and lack of redundancy, translating business rules and functionality requirements into ETL procedures.
* Wrote Korn shell scripts to load data to target and for running mappings.
* Developed almost all transformations such as Filter, Aggregator, Expression, Router, Lookup, Joiner, Update Strategy and Rank.
* Extensively involved in Performance Tuning issues.
* Involved in creating tasks, worklets, workflows and schedule, run, monitor sessions by using Workflow manager and workflow monitor.
* Migrated Mappings, Sessions and Workflows from Development to Test and then to Pre- Production environment.
* Tested the data and data integrity among various sources and targets. Involved in developing ad-hoc and business reports according to the requirement.

## Client : ABN AMRO, Netherland

##### Project : Remote Marketing Intelligence

## Role : Test Lead

## Tools : Informatica, Mainframe, UNIX shell scripting

## Duration : Sep 2012 to May 2013

## Location : Amsterdam, Netherland

Overview:

The project involves building of Informatica mappings to enhance quality of data delivered to customers. Major role is to test various applications developed by bank and providing resolution in case of defect(s) and implementation of the project.

*Responsibilities:*

* Tested the data and data integrity among various sources and targets. Involved in developing ad-hoc and business reports according to the requirement.
* Re-engineered on existing mappings to support new/changing business requirements.
* Used mapping, variables/parameters, and parameter files to support change data capture and workflow execution process.
* Modified several of the existing mappings based on the user requirements and maintained existing mappings, sessions and workflows.
* Used various transformations like Source Qualifier, Expression, Aggregator, Joiner, Filter, Lookup, and Update Strategy for Designing and optimizing the Mapping.
* Tuned the performance of mappings by following Informatica best practices and also applied several methods to get best performance by decreasing the run time of workflows.
* Prepared SQL Queries to validate the data in both source and target databases.
* Validated the Mappings, Sessions & Workflows, Generated & Loaded the Data into the target database.
* Monitored batches and sessions for weekly and Monthly extracts from various data sources to the target database.
* Exhaustive testing of applications using OPC scheduler (Mainframe).
* Monitoring batch run every week and providing resolutions in case of error.
* Creating change management to mention various components and plan to execute various tasks in production.

## Client : ABN AMRO, Netherland

##### Project : Remote Marketing Intelligence

## Role : Developer/Technical Lead

## Tools : Informatica, Mainframe, UNIX shell scripting

## Duration : Feb 2010 to Sep 2012

## Location : Amsterdam, Netherland

## Mumbai, India

## Bangalore, India

*Overview:*

**Remote Marketing Intelligence** architecture used in ABN AMRO is designed to support the rapid development and efficient execution of ETL applications using Informatica. It involves maintenance and organization of marketing database as present on Teradata. The project also implements best-practice methods of code version control and code migration for applications.

*Responsibilities:*

* Creating change management to mention various components and plan to execute various tasks in production.
* Involved in development of business rule documents and providing suggestions to functional business analysts to implement requirement.
* Development and design of code using Informatica and Teradata as development tools as well as performance improvement of existing applications.
* Exhaustive testing of applications using mainframe and SQL scripts in Teradata.
* Onsite offshore co-ordination/team handling from Offshore.
* **ONSITE** working experience that involves project implementation and requirement gathering.
* Solely responsible in successful implementation of some major projects.

# Certification and Training:

# Certification in course from “Data analyst training” and “Developer training for Apache Spark and Hadoop” courses from Cloudera.

* Exhaustive training on Informatica, Ab initio.
* Certification in Python Programming Course from University of Michigan.
* Certification in AWS certified cloud practitioner course from Linux Academy.