

Lead Python/Spark/Redshift/S3 Lead Python/Spark/Redshift/S3 Lead Python/Spark/Hadoop/Neo4j  
Charlotte, NC More than 12 years of experience in Python, Hive, Sqoop, Spark, Pig Scripts, Big Data, Neo4j, Neo4j Bloom, MS SQL Server, MSBI (SSIS, SSRS, SSAS), ASP.Net with C#.

Implemented Hadoop stack and different big data analytic tools, migration from different databases (i.e. Teradata, Oracle, MYSQL) to Hadoop. Experience in working with MapReduce programs using Hadoop for working with Big Data. Strong Knowledge of Hadoop and Hive and Hive's analytical functions. Capturing data from existing databases that provide SQL interfaces using Sqoop. Efficient in building hive and pig scripts. In depth understanding of Spark Architecture including Spark Core, Spark Sql, Data Frames. Experience in usage of Hadoop Distribution like Cloudera 5.3 Expertise in using Spark-SQL with various data sources like JSON and Parquet. Having good experience in Hadoop Big Data processing. Expertise in developing the queries in Hive, Pig. Successfully loaded files to Hive and HDFS from Mysql, Oracle and Teradata. Experience in designing both time driven and data driven automated workflows using Oozie. Mastering/Leading in the development of applications/tools using Python. Worked on several python packages like NumPy, SciPy, Pandas etc. Having Good Experience in Object Oriented Concepts with Python. Integrated different data sources, data wrangling: cleaning, transforming, merging and reshaping data sets by writing Python scripts. Designed and implemented Data Lineage graph from origination to consumption point of view in the form of nodes and edges in Neo4j. Created Nodes and Edges in NoSQL Database Neo4j. Strong Experience in implementing Data warehouse solutions in Confidential Redshift, migrate data from on premise databases to Confidential Redshift, RDS and S3. Expertise in writing Spark RDD transformations, actions, Data Frame's, case classes for the required input data and performed the data transformations using PySpark. Expertise in using Spark-SQL with various data sources like JSON and Hive. Expert in SSIS Package development, deployment, Job Scheduling and Error Trapping, SSRS Report development, deployment, subscription, SSAS Tabular and Multidimensional Cube development. Strong experience as Business Intelligence Developer and Data Analyst in Production, Development and Staging Environments. Defining data warehouse (star and snow

flake schema), fact table, cubes, dimensions, measures using SQL Server Analysis Services with DAX. Authorized to work in the US for any employer Work Experience Lead Python/Spark/Redshift/S3 Johnson & Johnson, New Jersey October 2018 to February 2019 Project Description: Goal of this project is to create data lake for the marketing and retail data from the multiple pharma vendors. This include retrieving huge datasets from vendors, create extracts on S3, validate that data using python scripts and migrate data into redshift and then create graph of customers( Practioners, Facility) and their relationships(Media Plan and Digital Media) Responsibilities: Designing and building full end-to-end Data Warehouse infrastructure from the ground up on Confidential Redshift for large scale data handling. Developed Spark program using PySpark API with Hive and SQL. Responsible for ETL and data validation using Python scripts. Worked on Quality Check script for matching extract count and trigger error in case of mismatch count in extracts using python. Worked on AWS Data Pipeline to configure data loads from S3 to into Redshift. Write complex stored procedures to extract data in Redshift. Import data from AWS S3 into Spark's RDD and performed transformations and actions on top of that RDD. Used AVRO, Parquet and ORC file format to store data into HDFS. Created Practioners and Facility as nodes and their relationships in neo4j. Used Apoc procedures to analyze data. Worked on boundary queries for media plan and digital media cypher queries. Deploy and execute python scripts using Unix command line. Environment: Python, PySpark Hive, HDFS, AWS, Redshift,Postgres, Mysql, S3, Neo4j 3.4 Lead Python/Hadoop/Neo4j AIG - Charlotte, NC October 2017 to September 2018 Project Description: Data Management Subsystems(DMS) consists of managing different systems workflows, processes, sub processes, applications and servers. Source for this subsystem is sql server and hive. Data is then integrated and validated by the python and then create nodes and relationships in neo4j graphical database. Workflows can be modified through java UI. Responsibilities: Worked on python files to load the data from csv, json, MySQL, hive files to Neo4j Graphical database. Implemented list and dictionary as data structure to manipulate data from sql server, csv and hive files. Manipulate dataframes using Pandas and Numpy Libraries in Python. Performed transformations and actions on PySpark's RDD. Import

csv, pyodbc and pyhive to deal with different structured files. Import data using Sqoop into Hive and from existing Teradata. Export and Import data into HDFS and Hive using Sqoop. Used JSON and XML Serde's for serialization and de-serialization to load Json and Xml data into Hive tables. Worked with AVRO, Parquet and ORC file format and various compression formats like Snappy to store data into HDFS. Load the data into Spark Dataframe and do in memory data computation to generate output response. Store data on HBASE for analysis. Wrote different pig scripts to clean up ingested data. Load and transform large sets of structured and semi structured data. Used sets to compare metadata and differentiate. Customized error logging as well as exceptional handling. Validate the extracted data and prepare valid files as well as data quality error files. Designed and implemented Data Lineage Neo4j graph from origination to consumption point of view. Developed Business process hierarchy graph for every process available in the system. Created indexes and constraints on nodes and edges of Neo4j Graph. Worked on boundary cypher queries using Apoc libraries. Create perspectives and parameterized queries in Neo4j Bloom. Extensive use of Apoc libraries for data modification and data analysis in Neo4j Graph. Design Neo4j architecture for nodes and edges load. Created Neo4j Graphical database nodes and relationship. Implement cypher queries to manipulate data on Neo4j database. Designed and developed a decision tree application using Neo4J graph database to model the nodes and relationships for each decision. Implemented data wrangling: cleaning, transforming, merging and reshaping data frames. Configure Neo4j on Unix as well as windows system.

Environment: Python, Hive, HBASE, Spark Sqoop, Oozie,PySpark,Pig Scripts, Teradata Neo4j 3.4,Neo4j Bloom, Sql Server 2017 Lead Python/Hadoop/Neo4j Otsuka,Los Angles - CA May 2017 to October 2017 Project Description: The project goal is to analyze pharma business by combining data from different sources (i.e. csv, excel, MySQL, hive, json, postgres, oracle) to one centralize place and then create nodes and relationships in neo4j graphical database. Responsibilities: Developed a Machine Learning application to predict disease medication based on demographic and medical history in python. Explored different implementations in Hadoop environment for data extraction and summarization by using packages like Hive, Pig. Importing and exporting data into

HDFS using Sqoop Sql Server, Teradata and Mysql. Manipulate dataframes using Pandas and Numpy Libraries in Python. Designed and developed a decision tree application using Neo4j graph database to model the nodes and relationships for each decision. Wrote python scripts to extract data from diverse sources. Worked on python scripts to load the data from csv, json, mysql, hive files to Neo4j Graphical database. Created Neo4j Graphical database nodes and edges using neo4j.v1 library in Python. Created indexes and constraints in Neo4j. Implement cypher boundary queries to manipulate data on Neo4j database. Data integration and data analysis using APOC procedures and functions in Neo4j. Involved in Design, analysis, Implementation, Testing and support of ETL processes for Stage, ODS and Mart. Used Talend as a Data Cleansing tool to correct the data before loading into the staging area. Collect and link metadata from diverse sources, including relational databases and flat files. Extracted data from different Flat files, MS Excel, Hlve and transformed the data based on user requirement using Talend and loaded data into target, by scheduling the sessions. Supporting daily loads and work with business users to handle rejected data. Implemented data cleansing for files using Talend. Created data model in Postgres using dimensional model. Performed Unit Testing and tuned for better performance. Created Reusable Transformations and multiple Mappings. Environment: Python, Hive, Pig, Sqoop, Oozie, Neo4j 3.2, Talend 9.6, PGAdmin 1.4 Data Engineer Florida Power and Light - Noida, Uttar Pradesh July 2016 to April 2017 Project Description: The System will streamline the verification of the sections like revenues billed, cash postings, refunds issued, meters uploaded and cash deposited to banks. The System has ability to create dashboard, custom reporting and analytics. The Revenue balance data can be drilled down by measures and dimensions, enabling users to gain access to accurate, up-to-date Information for better decision making. Responsibilities: Developed and maintained Python ETL scripts to scrape data from external sources and load cleansed data into a Sql Server. The data was used for daily electrical power virtual trading activities in several markets. Implemented discretization and binning, data wrangling: cleaning, transforming, merging and reshaping data frames using Python. Integrated applications with designing database architecture and server scripting, studying & establishing

Designed data visualization to present current impact and growth. Loaded the dataset into Hive for ETL Operation. Written Hive scripts to extract data from staging tables. Building, publishing customized interactive reports and dashboards, report scheduling using Power BI Desktop. Used Power BI Power Pivot to develop data analysis prototype and used Power View and Power Map to visualize reports. Used DAX (Data Analysis Expressions) functions for the creation of calculations and measures in the Power BI Models. Designing and apply Microsoft Project Plan. Requirement, Design, Development and Testing. Data Modeling and Cube design in SSAS using Tabular Model. Monitoring overall progress and use of resources, initiating corrective action where necessary. Planning and monitoring the project. Preparing and maintaining project, stage and exception plans as required. Environment: Hadoop, Hive, Python, SSAS, Power BI Data Engineer HMS, Irvine - Noida, Uttar Pradesh March 2014 to July 2016 Project Description: Client is an innovative healthcare services company with a Comprehensive, patient - centered approach to diabetes management. SSIS has been used as an ETL tool for building of data warehouse and data consolidation. Client provides the claim, provider and membership files on daily, weekly and monthly basis on FTP or EDI Server. Package load the claim data into staging database. After loading Medicare and Medicaid data in staging edits (Rules provided by State and Federal government for claims) applied on staging database and then after finally loaded into the Data Warehouse and then mail delivery report to the client at every step as Sanity Check, Return file and Cleanup process. Every step maintains in Sql server log table. Error handling and monitoring the package is done by the BixPress tools. Responsibilities: Designing, developing, deploying, job scheduling reports and success/failure notifications in MS SQL Server environment using SSIS in Sql Server data Tools (SSDT). Used ETL (SSIS) tasks to develop jobs for extracting, cleaning, transforming and loading data into data warehouse. Extracted data from various sources like SQL Server 2008, Oracle, CSV, Excel and Text file from Client servers and through FTP. Experience working in a large data environment doing SQL development. SQL query optimization (Knowledge of Transactions, Isolation Levels, Indexes, Blocks, Locks, and Database Partitioning etc.) Implemented SQL constraints (Primary Key, Foreign Key, Index, Unique, Not Null, Default) Experience managing

large projects with a close attention to detail and high-quality performance. Data modelling and mapping using UML in Visio. Experience in handling Enterprise datasets. Identify issues and fix those issues while troubleshooting. Experience in writing MDX queries to access cubes data in SSAS. Monitor scheduled SSIS Package through the BlxPress tools Error Trapping through BlxPress tool, Script Writing in .net, Transaction, Email Alert and some complex Transformation in SSIS Deployment and Scheduling SSIS Package through SQL Server and file system Report development in SSRS Create SSIS packages, Email alert for SSIS packages and schedule it through SQL Server and file system. Manage the Dev, SQA and Production Environments.

Environment: SQL Server 2012 with Integration, Reporting Services and Analysis Services, C#.net code for SSIS scripting, BlxPress, Microsoft Business intelligence development studio, SQL Server 2012. Data Engineer Sanare - Miami, FL September 2012 to March 2014 Project Description: The system is a database driven online Reporting model and SSIS package for data upload and send reports by the packages, which is for Health care domain in USA, for this report we have create our own data warehouse and fetch all report from this data warehouse. Around 100+ business reports have been created using functionality of SSRS. SSIS has been used as an ETL tool for building of data warehouse and data consolidation. ETL process has been designed on following lines

Download daily extracts from SFTP server and upload them into data warehouse after performing various data integrity checks Preparation of data comparison sheet to compare record counts in tables against the counts which are supposed to be them Execution of data maintenance tasks e.g. Rebuild Indices After successful completion of data upload this package kicks off second package to generate various reports and distribute them in encrypted form to intended recipients Preparation of logs has been maintained during execution of package Notifications to users of failure of package, if occurred Notifications to users of success of important tasks to keep them aware during execution At last summary has been sent to show the statistics of package Real time data has been picked from Other production systems for integrated reporting and Dashboards Package is configurable, and data driven so it can be changed easily as per the changing business environment. Responsibilities: Data warehouse architecture designing. ETL package

Development in SSIS. Deployment and Scheduling of SSIS Package in SSIS. Populate data marts with the right data at the right time. Performance Tuning/ SQL Query Optimization, Complex Query Design. Stored Procedure Writing, Function Writing. Creating Reports in SSRS. Project monitoring, Status reporting and client interaction. Monitoring the critical components of the System. Implementation and support of systems in production environment. Environment: SQL Server 2008 R2 with Integration and Reporting Services, SSIS Scripting, Microsoft Business intelligence development studio, SQL Server 2008 R2 ETL Developer Aspen - Noida, Uttar Pradesh April 2012 to September 2012 Project Description: Data Repository collects all clinical, operational, and financial extracts across the organization and uploads them in data warehouse after performing various data integrity checks. It also integrates real time cloud data with the ware house and provides an open environment for powerful reporting and decision support to healthcare managerial authorities. SSIS has been used as an ETL tool with following features: Created Complex ETL Packages using SSIS to extract data from staging tables to partitioned tables with incremental load.

Created SSIS Reusable Packages to extract data from Multi formatted Flat files, Excel, XML files into Database Billing Systems. Created SSIS packages for File Transfer from one location to the other using FTP task. Developed, deployed, and monitored SSIS Packages Transform extract data after performing integrity checks. Implement Email notification in case of package failure. Integration with another package for clean-up task. Develop various script tasks for validating and transform data. Designed SSIS Packages to transfer data from flat files to SQL Server using Business Intelligence Development Studio. Extensively used SSIS transformations such as Lookup, Derived column, Data conversion, Aggregate, Conditional split, SQL task, Script task and Send Mail task etc. Implement configuration files to deploy the SSIS packages across all environments. Responsibilities: ETL package Development in SSIS. Integration of real-time cloud data (NetSuite) with data Warehouse for critical business requirements. Report development in SSRS. Project monitoring, Status reporting and client interaction. Operate as the liaison between Digital Analytics team and IT Development team for all projects and releases including testing pre/post deployment of changes to ensure timely and accurate reporting and insights

Environment: SQL Server 2008 R2 with Integration and Reporting Services, Microsoft Business intelligence development studio, SSMS, SQL Server 2008 R2 ETL and Report Developer Aspen - Noida, Uttar Pradesh September 2011 to March 2012 Project Description: PMIS is an application which stores information about various types of Patients disease and taking input from different chain management solutions around the globe. PMIS is used to manage patient historical information. Various business reports have been developed for dynamic data analysis. Comparison of actual health data with industry standard in graphical analysis has been shown. Following features have been implemented in this project: Designing, developing and deploying reports in MS SQL Server environment using SSRS-2008 and SSIS in Business Intelligence Development Studio (BIDS). Used ETL (SSIS) to develop jobs for extracting, cleaning, transforming and loading data into data warehouse. Designed SSIS Packages to transfer data from flat files to SQL Server using Business Intelligence Development Studio. Creating multiple parameterized stored procedures which were used by the reports to get the data. Worked on formatting SSRS reports using the Global variables and expressions. Created parameterized reports, Drill down and Drill through reports using SSRS. Used Execution Plan, SQL Profiler and Database Engine Tuning Advisor to optimize queries and enhance the performance of databases. Optimized query performance by creating indexes. Write T-SQL statements for retrieval of data and performance tuning of TSQL. Responsibilities: ETL package Development in SSIS. Report development in SSRS. Database Design. Understanding the business problem, identifying relevant data, gathering and summarizing data meaningfully. Performance Tuning/Query Optimization in SPs. Complex Query Design, Stored Procedure Writing. Function Writing, Trigger Writing.

Environment: SQL Server 2008 R2/SSIS/SSRS, Microsoft Business intelligence development studio, SQL Server 2008 R2 Asp.Net Developer Oil Field Services - Mumbai, Maharashtra September 2010 to September 2011 Project Description: OFS is web-based application that keeps track of products between stores/rigs/well. OFS application offers clients a secure and efficient management tool in processing their business. OFS application is built as an n-tier Web based application. The Application User interface is built as Web application using ASP.NET Master



Pages, AJAX controls and JavaScript. To Support SOA based architecture, Business Web services in middle tier is used and it acts as core business engine. The web services consume/ provide the Business Entities. Business Entities are used for passing the data across the tiers. Database tier is the oracle database. To maintain application portability across databases, data transactions are handled in business services tier. Responsibilities: Coordination and handling tasks among team members. Designing of HTML pages, JavaScript validations. Assist Project Manager in Feasibility study and requirement Analysis phase. Implement web service as business layer. Deploy application on Client Server. Designing Database, Creation and implementation of View, Stored Procedures. Documentation for coding. Configuring, deploying Database and web application for testing environment. Code review during different stage of development life cycle.

Environment: Web, Framework 2.0, ASP.Net 2.0, C#, Web Service, TFS, Visual Studio 2005, JavaScript, Oracle 10G Asp.Net Developer Insurity, Hartford - Mumbai, Maharashtra January 2010 to August 2010 Project Description: NGA (Next Generation Application) is a web-based insurance-product of one of the Famous American product-based company in commercial lines. It allows their clients to do policy-administration as, Automated rating and issuance support for all major commercial lines of business. Full policy lifecycle transaction support Full maintenance of ISO and customer-specific rates, forms, and rules changes. Responsibilities: Requirement Gathering of defects. Onsite Coordination. Estimation of defects. Defect Fixing Code review

Environment: Web, Framework 2.0, VB.Net, VB6, SQL Server 2005 Asp.Net Developer Insurity, Hartford - Mumbai, Maharashtra March 2009 to January 2010 Project: CHealth Project Description: CHealth is a Collaborative Consumer Engagement, which is based on Microsoft connected Healthcare framework. It empowers consumers through its Wellness module and at the same time provides an integrated platform wherein Payers, TPAs, Brokers, ASOs & Providers collaborate seamlessly. With an innovative business model cHealth is supported by SOA based architecture. To Support SOA based architecture, Business Web services in middle tier is used and it acts as core business engine. Responsibilities: Coordination and handling tasks among team members. Designing of HTML pages. Assist Project Manager in Feasibility study and requirement Analysis

phase. Deploy application on Client Server. Database Design, Creation and implementation of View, Stored Procedures. Taking Backup and Restore Database. Implement Web services as business tier. Documentation for coding. Configuring, deploying Database and web application for testing environment. Code review during different stage of development life cycle. Developed HL7 translator. Write Store Procedure to implement HL7 translator. Environment: Web, Framework 2.0, Asp.Net with C#, SQL Server 2005 Asp.Net Developer Munich Re, Princeton - Mumbai, Maharashtra August 2007 to February 2009 Project Description: AutoFac is a web-based application that allows Re-Insurance clients to conduct Property and Commercial Facultative re-insurance business operations with MRAM. AutoFac application offers clients a secure and efficient management tool in processing their business. AutoFac application is built as an n-tier Web based application. The Application User interface is built as Web application using ASP.NET Master Pages, AJAX controls and JavaScript. To Support SOA based architecture, Business Web services in middle tier is used and it acts as core business engine. The web services consume/ provide the Business Entities. Business Entities are used for passing the data across the tiers. Database tier is the oracle database. To maintain application portability across databases, data transactions are handled in business services tier. Responsibilities: Designing of HTML pages, JavaScript validations. Programming for Referral, People, and System admin, Organization, Policy, Quotes Location Upload and Reporting. Identifying the Functionality of the Application. Designing of framework. Designing Database Model, Creating Views and Stored Procedure. Requirement gathering phase and UAT. Extensive work on web services as business tier. Deploying application for testing environment. Documentation for specific modules Integration and System Testing Programming in XSLT for location upload Environment: Web, Framework 2.0, Microsoft Framework 2.0(ASP.Net) with C#, Oracle 10G Education Master's Skills APACHE HADOOP SQOOP (1 year), ASP (4 years), Asp.Net (4 years), Business intelligence (5 years), C# (5 years), database. (10+ years), ETL (6 years), EXTRACT, TRANSFORM, AND LOAD (6 years), Hadoop (2 years), Hive (2 years), HTML (3 years), JavaScript. (2 years), MS ASP (4 years), MS SQL Server (8 years), Python (2 years), R2 (2 years), SAP (2 years), SQL. (8 years), sql server (8 years), SQL

Server 2005 (1 year)

Name: Jennifer Morgan

Email: drodriguez@example.com

Phone: (354)505-5560x56579