Name : XXXXXXXXXXXXXXXX
Ph.No: XXXXXXXXXXXXX
Email : XXXXXXXXXXXXXXXXX

Summary:

- 8+ years of extensive experience in Information Technology wif expertise on Data Analytics, Data Architect, Design, Development, Implementation, Testing and Deployment of Software Applications in Banking, Finance, Insurance, Retail and Telecom domains.
- Working experience on designing and implementation complete end to end Hadoop infrastructure using HDFS, MapReduce, Hive, HBase, Kafka, Sqoop, Spark, zookeeper, Ambari, Scala, Oozie, Yarn, No SQL, Postman and Python
- Created Data Frames and performed analysis using Spark SQL.
- Acute noledge on Spark Streaming and Spark Machine Learning Libraries.
- Hands on expertise in writing different RDD (Resilient Distributed Datasets) transformations and actions using Scala, Python and Java.
- Excellent understanding of Spark Architecture and framework, Spark Context, APIs, RDDs, Spark SQL, Data frames, Streaming, MLlib.
- Worked in agile projects delivering end to end continuous integration/continuous delivery pipeline by Integration of tools like Jenkins and AWS for VM provisioning.
- Experienced in writing teh automatic scripts for monitoring teh file systems, key MapR services.
- Implemented continuous integration & deployment (CICD) through Jenkins for Hadoop jobs.
- Good Knowledge on Cloudera distributions and in Amazon simple storage service (Amazon S3), AWS Redshift, Lambda and Amazon EC2, Amazon EMR.
- Excellent understanding of Hadoop Architecture and good Exposure in Hadoop components like Hadoop Map Reduce, HDFS, HBase, Hive, Sqoop, Cassandra, Kafka and Amazon Web services (AWS) API test, document and monitor by Postman which is easily integrate teh tests into your build automation.
- Used Sqoop to Import data from Relational Database (RDBMS) into HDFS and Hive, storing using different formats like Text, Avro, Parquet, Sequence File, ORC File along wif compression codes like Snappy and Gzip.
- Performed transformations on teh imported data and Exported back to RDBMS.
- Worked on Amazon Web service (AWS) to integrate EMR wif Spark 2 and S3 storage and Snowflake.
- Experience in writing queries in HQL (Hive Query Language), to perform data analysis.
- Created Hive External and Managed Tables.
- Implemented Partitioning and Bucketing on Hive tables for Hive Query Optimization.
- Used Apache Flume to ingest data from different sources to sinks like Avro, HDFS.
- Implemented custom interceptors for flume to filter data and defined channel selectors to multiplex teh data into different sinks.
- Excellent noledge on Kafka Architecture.
- Integrated Flume wif Kafka, using Flume both as a producer and consumer (concept of FLAFKA).
- Used Kafka for activity tracking and Log aggregation.
- Experienced in writing Oozie workflows and coordinator jobs to schedule sequential Hadoop jobs.
- Experience working wif Text, Sequence files, XML, Parquet, JSON, ORC, AVRO file formats and Click Stream log files.
- Familiar in data architecture including data ingestion pipeline design, Hadoop architecture, data modeling and data mining and advanced data processing. Experience optimizing ETL workflows.
- Good Exposure in Data Quality, Data Mapping, Data Filtration using Data warehouse ETL tools like Talend, Informatica, Data Stage, Ab - initio
- Good Exposure to create various dashboard in Reporting Tools like SAS, Tableau, Power BI, BO, QlikView used various filters, sets while dealing wif huge volume of data.
- Experience in various Database such as Oracle, Teradata, Informix and DB2.
- Experience wif NoSQL like MongoDB, HBase and PostgreSQL like Greenplum
- Worked in complete Software Development Life Cycle like Analysis, Design, Development, Testing, Implementation and Support using Agile and Waterfall Methodologies.
- Demonstrated a full understanding of teh Fact/Dimension data warehouse design model, including star and snowflake design methods.

Technical Skills:

- **Big Data Ecosystem:** HDFS, MapReduce, HBase, Pig, Hive, Sqoop, KafkaFlume, Cassandra, Impala, Oozie, Zookeeper, MapR, Amazon Web Services (AWS), EMR
- **Machine Learning:** Classification AlgorithmsLogistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Gradient Boosting Classifier, Extreme Gradient Boosting Classifier, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes Classifier, Extra Trees Classifier, Stochastic Gradient Descent, etc.
- **Cloud Technologies:** AWS, Azure, Google cloud platform (GCP)
- **IDE's IntelliJ:** Eclipse, Spyder, Jupyter
- **Ensemble and Stacking:** Averaged Ensembles Weighted Averaging, Base Learning, Meta Learning, Majority Voting, Stacked Ensemble, AutoML - Scikit-Learn, MLjar, etc.
- **Databases:** Oracle 11g/10g/9i, MySQL, DB2, MS SQL Server, HBASE
- **Programming:** Query LanguagesJava, SQL, Python Programming (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib, NLTK), NoSQL, PySpark, PySpark SQL, SAS, R Programming (Caret, Glmnet, XGBoost, rpart, ggplot2, sqldf), RStudio, PL/SQL, Linux shell scripts, Scala.
- **Data Engineer:** Big Data Tools / Cloud / Visualization / Other ToolsDatabricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, Mahout, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Salesforce, GCP, Google Shell, Linux, PuTTY, Bash Shell, Unix, etc., Tableau, Power BI, SAS, We Intelligence, Crystal Reports, Dashboard Design.
- **Versioning tools:** SVN, Git, GitHub
- **Operating Systems:** Windows 7/8/XP/2008/2012, Ubuntu Linux, MacOS
- **Network Security:** Kerberos
- **Database Modelling:** Dimension Modeling, ER Modeling, Star Schema Modeling, Snowflake Modeling
- **Monitoring Tool:** Apache Airflow
- **Visualization/ Reporting:** Tableau, ggplot2, matplotlib, SSRS and Power BI
- **Machine Learning Techniques:** Linear & Logistic Regression, Classification and Regression Trees, Random Forest, Associative rules, NLP and Clustering.
- **Build and CI tools:** Docker, Kubernetes, Maven, Gradle, Jenkins, Hudson, Bamboo
- **SDLC Methodologies:** Agile, Waterfall**,** Scrum, TDD

Work Experience:

T D Bank, Bellevue, Washington **Jan 2022 –Till date**

Sr. GCP Data Engineer

Responsibilities:

- Migrating an entire oracle database to BigQuery and using of power bi for reporting.
- Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
- Experience in GCP Dataproc, GCS, Cloud functions, BigQuery.
- Experience in moving data between GCP and Azure using Azure Data Factory.
- Experience in building power bi reports on Azure Analysis services for better performance.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
- Coordinated with team and Developed framework to generate Daily adhoc reports and Extracts from enterprise data from BigQuery.
- Experience architecting and implementing data visualizations that visually present complex data and/or metric relationships to users for exploration, analysis and action
- Experience building wireframes, mock-ups, and ad-hoc visualizations/web pages that present the "art of the possible" for users to interact with and provide direction
- Designed and Co-ordinated with Data Science team in implementing Advanced Analytical Models in Hadoop Cluster over large Datasets.
- Wrote scripts in Hive SQL for creating complex tables with high performance metrics like partitioning, clustering and skewing
- Work related to downloading BigQuery data into pandas or Spark data frames for advanced ETL capabilities.

- Worked with google data catalog and other google cloud API's for monitoring, query and billing related analysis for BigQuery usage.
- Worked on creating POC for utilizing the ML models and Cloud ML for table Quality Analysis for the batch process.
- Knowledge about cloud dataflow and Apache beam.
- Good knowledge in using cloud shell for various tasks and deploying services.
- Created BigQuery authorized views for row level security or exposing the data to other teams.
- Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, SQOOP, Apache Spark, with Cloudera Distribution.

Environment: GCP, AWS, JMeter, Kafka, Ansible, Jenkins, Docker, Maven, Linux, Red Hat, GIT, Cloud Watch, Python, Shell Scripting, Golang, Web Sphere, Splunk, Tomcat, Soap UI, Kubernetes, Terraform, PowerShell

T-Mobile, Cherry hills, NJ **Oct 2019 - Jan 2022**
Sr. AWS Data Engineer
Responsibilities:

- Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services (AWS) on EC2.
- Handled AWS Management Tools as Cloud watch and Cloud Trail.
- Stored teh log files in AWS S3. Used versioning in S3 buckets where teh highly sensitive information is stored.
- Integrated AWS Dynamo DB using AWS lambda to store teh values of items and backup teh DynamoDB streams
- Automated Regular AWS tasks like snapshots creation using Python scripts.
- Designed data warehouses on platforms such as AWS Redshift, Azure SQL Data Warehouse, and other high-performance platforms.
- Install and configure Apache Airflow for AWS S3 bucket and created dags to run teh Airflow
- Prepared scripts to automate teh ingestion process using Pyspark and Scala as needed through various sources such as API, AWS S3, Teradata and Redshift.
- Created multiple scripts to automate ETL/ ELT process using Pyspark from multiple sources
- Developed Pyspark scripts utilizing SQL and RDD in spark for data analysis and storing back into S3
- Developed Pyspark code to load from stg to hub implementing teh business logic.
- Developed code in Spark SQL for implementing Business logic wif python as programming language.
- Designed, Developed and Delivered teh jobs and transformations over teh data to enrich teh data and progressively elevate for consuming in teh Pub layer of teh data lake.
- Worked on Sequence files, Map side joins, bucketing, partitioning for hive performance enhancement and storage improvement.
- Wrote, compiled, and executed programs as necessary using Apache Spark in Scala to perform ETL jobs wif ingested data.
- Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
- Maintained Kubernetes patches and upgrades.
- Managed multiple Kubernetes clusters in a production environment.
- Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation and used Spark engine, Spark SQL for data analysis and provided to teh data scientists for further analysis
- Developed various UDFs in Map-Reduce and Python for Pig and Hive.
- Data Integrity checks has been handled using hive queries, Hadoop, and Spark.
- Worked on performing transformations & actions on RDDs and Spark Streaming data wif Scala.
- Implemented teh Machine learning algorithms using Spark wif Python.
- Profile structured, unstructured, and semi-structured data across various sources to identify patterns in data and Implement data quality metrics using necessary query's or python scripts based on source.
- Designs and implementing Scala programs using Spark Data frames and RDDs for transformations and actions on input data.
- Improved teh Hive queries performance by implementing partitioning and clustering and Optimized file formats (ORC).

Environment: AWS, JMeter, Kafka, Ansible, Jenkins, Docker, Maven, Linux, Red Hat, GIT, Cloud Watch, Python, Shell Scripting, Golang, Web Sphere, Splunk, Tomcat, Soap UI, Kubernetes, Terraform, PowerShell.

Humana, Louisville, KY **Aug 2018 - Sep 2019**

Big Data Engineer & AWS Cloud Engineer

Responsibilities:

- Worked on AWS Data pipeline to configure data loads from S3 to into Redshift.
- Using AWS Redshift, me Extracted, transformed and loaded data from various heterogeneous data sources and destinations.
- Created Tables, Stored Procedures, and extracted data using T-SQL for business users whenever required.
- Performs data analysis and design, and creates and maintains large, complex logical and physical data models, and metadata repositories using ERWIN and MB MDR
- wrote shell script to trigger data Stage jobs.
- Assist service developers in finding relevant content in teh existing reference models.
- Like Access, Excel, CSV, Oracle, flat files using connectors, tasks and transformations provided by AWS Data Pipeline.
- Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
- Worked on developing Pyspark script to encrypting teh raw data by using Hashing algorithms concepts on client specified columns.
- Responsible for Design, Development, and testing of teh database and Developed Stored Procedures, Views, and Triggers
- Created Tableau reports wif complex calculations and worked on Ad-hoc reporting using PowerBI.
- Creating datamodel dat correlates all teh metrics and gives a valuable output.
- Worked on teh tuning of SQL Queries to bring down run time by working on Indexes and Execution Plan.
- Exploring wif Spark to improve teh performance and optimization of teh existing algorithms in Hadoop using Spark context, Spark-SQL, postgreSQL, Data Frame, OpenShift, Talend, pair RDD's
- Involved in integration of Hadoop cluster wif spark engine to perform BATCH and GRAPHX operations.
- Performed data preprocessing and feature engineering for further predictive analytics using Python Pandas.
- Generated report on predictive analytics using Python and Tableau including visualizing model performance and prediction results.
- Implemented Copy activity, Custom Azure Data Factory Pipeline Activities
- Primarily involved in Data Migration using SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.
- Implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB).
- Migration of on premise data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake and Stored (ADLS) using Azure Data Factory (ADF V1/V2).
- Developed a detailed project plan and helped manage teh data conversion migration from teh legacy system to teh target snowflake database.
- Design, develop, and test dimensional datamodels using Star andSnowflakeschemamethodologies under teh Kimball method.
- Implement ad-hoc analysis solutions using Azure Data Lake Analytics/Store, HDInsight
- Developed data pipeline using Spark, Hive, Pig, python, Impala, and HBase to ingest customer
- Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
- Worked on a direct query using PowerBI to compare legacy data wif teh current data and generated reports and stored and dashboards.
- Designed SSIS Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for teh SSAS cubes (OLAP) SQL Server reporting services (SSRS). Created & formatted Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Subreports, ad-hoc reports, parameterized reports, interactive reports & custom reports
- Created action filters, parameters and calculated sets for preparing dashboards and worksheets using PowerBI.
- Developed visualizations and dashboards using PowerBI
- Sticking to ANSI SQL language specification wherever possible, and providing context about similar functionality in other industry-standard engines (e.g. referencing PostgreSQL function documentation)
- Used ETL to implement teh Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
- Performing ETL testing activities like running teh Jobs, extracting teh data using necessary queries from database transform, and upload into teh Data warehouse servers.

- Created dashboards for analyzing POS data using Power BI.

Environment: MS SQL Server 2016, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel (creating formulas, pivot tables, Hlookup, Vlookup, Macros), Spark, Python, ETL, Power BI, Tableau, Presto, Hive/Hadoop, Snowflakes, Power BI, AWS Data Pipeline, IBM Cognos 10.1, Data Stage, Cognos Report Studio 10.1, Cognos 8 & 10 BI, Cognos Connection, Cognos office Connection, Cognos 8.2/3/4, Data stage and Quality Stage 7.5

Independence Health Group, Bakersfield, CA **Mar 2016 – July 2018**
Hadoop Engineer/Data Engineer
Responsibilities:

- Involved in complete Implementation lifecycle, specialized in writing custom MapReduce, and Hive
- Extensively used Hive/HQL or Hive queries to query or search for a string in Hive tables in HDFS
- Continuous monitoring and managing teh Hadoop cluster using Cloudera Manager
- Implemented Spark using Python and Spark SQL for faster processing of data
- Used Spark for interactive queries, processing of streaming data and integration wif popular NoSQL database
- Used teh Spark -Cassandra Connector to load data to and from Cassandra
- Implemented test scripts to support test driven development and continuous integration.
- Dumped teh data from HDFS to Oracle database and vice-versa using Sqoop
- Extensively involved in Installation and configuration of Cloudera Hadoop Distribution.
- Provided support for EBS, Trusted Advisor, Cloud Watch, Cloud Front, IAM, Security Groups, Auto-Scaling, AWS CLI and Cloud Watch Monitoring creation and update.
- Worked wif Amazon Web Services (AWS) using EC2 for computing and S3 as storage mechanism
- Deployed Lambda and other dependencies into AWS to automate EMR Spin for Data Lake jobs
- Scheduled spark applications/Steps in AWS EMR cluster.
- Extensively used event-driven and scheduled AWS Lambda functions to trigger various AWS resources.
- Implemented advanced procedures like text analytics and processing using teh in-memory computing capabilities like Apache Spark written in Scala.
- Developed spark applications for performing large scale transformations and denormalization of relational datasets.
- Developed and executed a migration strategy to move Data Warehouse from SAP to AWS Redshift.
- Loaded data into teh cluster from dynamically generated files using Flume and from relational database management systems using Sqoop.
- Used Spark Streaming to divide streaming data into batches as an input to spark engine for batch processing.
- Worked on analyzing Hadoop cluster and different Big Data analytic tools including Pig, hive, HBase, Spark and Sqoop.
- Exported data from HDFS to RDBMS via Sqoop for Business Intelligence, visualization, and user report generation.
- Loading teh data from multiple Data sources like (SQL, DB2, and Oracle) into HDFS using Sqoop and load into Hive tables.
- Performed Real time event processing of data from multiple servers in teh organization using Apache Storm by integrating wif apache Kafka.
- Performed Impact Analysis of teh changes done to teh existing mappings and provided teh feedback
- Create mappings using reusable components like worklets, mapplets using other reusable transformations.
- Participated in providing teh project estimates for development team efforts for teh offshore as well as on-site.
- Coordinated and monitored teh project progress to ensure teh timely flow and complete delivery of teh project
- Worked on Informatica Source Analyzer, Mapping Designer & Mapplet, and Transformations.

Environment: Hadoop, HDFS, Hive, MapReduce, Impala, Sqoop, SQL, Informatica, Python, Flume, PySpark, Yarn, Pig, Oozie, Linux, AWS, Tableau, Maven, Jenkins, Cloudera, SAS (BI & DI), PL/SQL, Autosys, Oracle, Sql Server, No Sql, Teradata.

Morgan Stanley, New York, NY **Jul 2015 - Mar 2016**
Python Data Engineer
Responsibilities

- Developed Data pipelines using python for medical image pre-processing, Training and Testing.

- Developed Artificial Intelligence Platform which helps Data Scientist's to Train, Test and develop A.I. models on Amazon Sagemaker.
- Used Pandas, Opencv, Numpy, Seaborn, Tensorflow, Keras, Matplotlib, Sci-kit-learn, NLTK in Python for developing data pipelines and various machine learning algorithms.
- Design and engineer REST APIs and/or packages that abstract feature extraction and complex prediction/forecasting algorithms on time series data.
- Developed Python application for Google Analytics aggregation and reporting and used Django configuration to manage URLs and application parameters.
- Developed pre-processing pipelines for DICOM and NONDICOM Images.
- Developed and presented analytical insights on medical data, image data.
- Implement AWS Lambdas to drive real-time monitoring dashboards from system logs.
- Cleansing the data for normal distribution by applying various techniques like missing value treatment, outlier treatment, and hypothesis testing.
- Perform Data Cleaning, features scaling, features engineering using pandas and numpy packages in python.
- Create several types of data visualizations using Python and Tableau.
- Collected data needs and requirements by Interacting with the other departments.
- Worked on different data formats such as JSON, XML.
- Performed preliminary data analysis using descriptive statistics and handled anomalies such as removing duplicates and imputing missing values.
- Developed various graph methods to visualize and understand the data like Scatter plot, Pi-plot, bar charts, box-plot, and histograms.
- Involved in development of Web Services using REST API's for sending and getting data from the external interface in the JSON format.
- Configured EC2 instances and configured IAM users and roles and created S3 data pipe using Boto API to load data from internal data sources.
- Developed rest API's using python with flask framework and done the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
- Implemented Agile Methodology for building an internal application.
- Developed A.I machine learning algorithms like Classification, Regression, Deep Learning using python.
- Conducted statistical analysis on Healthcare data using python and various tools.
- Experience in cloud versioning technologies like Github.
- Worked closely with Data Scientists to know data requirements for the experiments.
- Deep experience in using DevOps technologies like Junkins,Docker, Kubernetes etc.

Alignment Healthcare, Orange, CA **Nov 2014 – Jun 2015**
Jr. Data Engineer
Responsibilities:

- Involved in defining teh source to target Data mappings, business rules and Data definitions.
- Responsible for Cluster maintenance, adding and removing cluster nodes, Cluster Monitoring and Troubleshooting, Manage and review data backups and log files.
- Loading teh data from multiple Data sources like (SQL, DB2, and Oracle) into HDFS using Sqoop and load into Hive tables.
- Built re-usable Hive UDF libraries which enabled various business analysts to use these UDF's in Hive querying.
- Performed data analysis in Hive by creating tables, loading it wif data and writing hive queries which will run internally in a MapReduce way.
- Responsible for running Hadoop MR jobs to process terabytes of xml's data.
- Used FLUME to export teh application server logs into HDFS.
- Installed Oozie workflow engine to run multiple Hive and MR jobs which run independently wif time and data availability.
- Used Spark to design and perform both batch processing (similar to MapReduce) and new workloads like streaming, interactive queries, and machine learning.

- Configured Spark streaming to receive real time data from teh Kafka and store teh stream data to HDFS using Scale.
- Implemented Spark using Scala and Spark SQL for faster testing and processing of data
- Design/Implement large scale pub-sub message queues using Apache Kafka.
- Developed Pig Latin scripts to extract teh data from teh web server output files to load into HDFS., NoSQL databases including HBase, MongoDB, and Cassandra.
- Designed HBase tables for time series data and Designed row key to avoid region server hot spotting.
- Used HBase API's to get and scan events data stored in HBase
- Deployed latest patches for Linux and Application servers, performed tuning.
- Involved in writing UNIX Shell and Perl scripts for automation of deployments to Application server.
- Experience on overall Hadoop architecture, data ingestion, data modelling and data mining.
- Handled importing data from different data sources into HDFS using Sqoop and performing transformations using Hive and then loading data into HDFS.
- Exporting of a result set from HIVE to MySQL using Sqoop export tool for further processing.
- Collecting and aggregating large amounts of log data and staging data in HDFS for further analysis.
- Experience in managing and reviewing Hadoop Log files.
- Used Sqoop to transfer data between relational databases and Hadoop.
- Worked on HDFS to store and access huge datasets wifin Hadoop.
- Good hands-on experience wif GitHub.

Environment: Cloudera Manager (CDH5), Hadoop, HDFS, Sqoop, Pig, Hive, Oozie, Kafka, flume, SQL Server, MySQL, Git.

Education:

- Bachelors in Computer Science from Osmania University in 2014.