**Professional Summary**

* Self-motivated Data scientist and machine learning engineer with **13+ years** of experience in identifying patterns in data analysis effectively with multi-functional roles to identify and leverage areas for improvement in data systems. bringing a comprehensive understanding of warehouse operations and documentation needs.
* **Data Science:** Data analysis, visualization, feature engineering, data mining, machine learning and data modeling.
* **Machine Learning:** Python, PySpark, and python libraries like NumPy, SciPy, sci-kit-learn, Pandas, TensorFlow, Kera’s, PyTorch and Matplotlib. In depth understanding of different supervised and unsupervised Machine Learning algorithms, recommendation systems and their implementation.
* **NLP:** Have hands on experience with libraries like NLTK, Text Blob, SpaCy, Gensim for text data cleaning, tokenization of data, normalization, chunking, POS tagging, language parsing, language quantification and have eloquent knowledge about machine learning and deep learning algorithms like BOW, tfidf, word embedding, word2vec, RNN models like LSTM, LLMs for sentiment analysis, chatbots and text generation.
* Strong experience in data Injection, data storage and data processing using Hadoop Ecosystem tools like **Sqoop**, **Hive, Pig, Spark, MapReduce, Spark Streaming, MapReduce, Flume, Kafka, HBase, Oozie, Zookeeper, and HDFS**.
* Good experience in **Scala, Python, Unix scripts and** writing **UDF’s** including SQL, JSON, XML.
* Excellent experience in importing and exporting data using **Sqoop from RDBMS to HDFS** and vice-versa.
* Well-versed in and implemented **Partitioning, Dynamic-Partitioning and bucketing** concepts in Hive to compute data metrics.
* Extensively worked with **SQL Server, HBase**, and **MySQL**.
* Excellent understanding and hands on experience using **NOSQL databases** like Mongo DB and **HBase**.
* Experience in developing data integration solutions in **Microsoft Azure Cloud Platform** using services **Azure Data Factory [ADF],** **Azure Synapse Analytics, Azure SQL Database [ADB], Azure Blob Storage, Azure Data Lake Storage [ADLS] and Azure DevOps**.
* Hands on experience troubleshooting, configuring Azure services like Azure Data factory, Azure Data Lake
* Experience in creating various datasets in **ADF** using linked services to connect to different source and target systems like **SQL Server, Oracle, Azure Blob Storage, Azure Data Lakehouse and Azure Synapse Analytics and Azure SQL DB**.
* Developed DataBricks ETL pipelines using notebooks with data frames and PySpark.
* Implemented various parameterized **Azure Data Factory pipelines** using activities like Copy activity, Custom **Azure Data Factory Pipeline** Activities.
* Self-hosted integration runtime has been created to copy the files form On-premises VM using various activities like metadata, foreach and copy and loaded into **ADLS GEN2** and azure synapse analytics.
* Developed the pipelines in **Azure Data factory** for various scenarios to meet business requirement using blob storages and ingesting the data into **azure synapse** analytics.
* Strong working in global delivery model (**Onsite-Offshore model**) involving multiple vendors and cross functional engineering teams.
* Excellent oral and written communication skills and great **team player.**

**Education**

* MS in Data Science, SMU University Jan’ 2018- Aug’ 2019
* MS in Computer Applications, Osmania University 2004 – 2007

**Certifications**

* NLP at CMU Jun’ 2023
* NLP specialist – Aug 2023 – December 2023A

**Accomplishments**

* Collaborated with a team of 6+ in the development of multiple projects.
* I make contributions to the team in developing data pipelines, data analytics, tuning and handling many production issues.
* I achieved Bravo/Recognition by completing tasks with accuracy and efficiency.

**Sr. DATA ENGINEER AND DATA SCIENTIST 08/2023 to 06/2024**

**BCBSM (Sr Health Services**) **Remote**

**Roles and Responsibilities:**

* The project is to identify the risk suspects and quality measures for the BCBSM population.
* Involved in assisting architect solution for this project and also participated in design.
* Lead and developed the effort for data ingestion framework, to ingest data from different sources like Oracle, MySQL.
* Once the data is in cloud storage blobs, used panda libraries to read the data frames and perform transformations.
* Extensively worked on analytics using Azure Databricks to implement SCD-1, SCD-2 approaches.
* Applied data mining technique to categorize the data based on SCD-1 and SCD-2 patterns.
* Developed Spark notebooks to transform and partition the data and organize files in ADLS Gen2.
* Machine learning algorithms are used to build predictive models that can forecast future risks based on historical claims, and drug usage data.
* Used unsupervised learning methods to segment the Medicare population according to use case and used those segments to create effective outreach strategy.
* Worked on creating Azure Blob for storing unstructured data in the cloud as blobs Data Modelling / Development of reports semantic layer on Azure Analysis services, Azure Synapses.
* The model goal is to identify and predict disease and provide personalized healthcare recommendations for providers. We used NumPy, Scikit Learn, Keras, TensorFlow and PyTorch libraries in Python Notebook for model development and predictions.
* Used predictive classification modeling, enabling healthcare providers to identify and target high-risk patients or those who would benefit the most from preventive care.
* Analyzed large claim and call datasets implemented (NLP – Topic Extraction) in python.
* Did ICD codes classification using sentimental analysis technique and categorize them based on risk measures.
* Did used MLflow for orchestrating machine learning algorithms and build ML pipeline.
* Used Airflow as a scheduler to run the workflows and schedule the jobs.
* The results are shared with business stakeholders for review and based on feedback work on it accordingly as per defined agile process.
* Being a Lead resource and having full knowledge of domain and data, played key role to business and tech team by mentoring team.
* **Environment:** Tools: Azure Data Factory, Airflow, Azure Databricks, Scala, Spark, Azure SQL DB, Azure DW (Synapse Analytics), Azure DevOps, Storage account, Python, Pandas, ScikitLearn, Numpy, Keras-Tensorflow, PyTorch, Airflow, ML flow

**Sr. DATA ENGINEER AND DATA SCIENTIST 09/2019 to 08/2023**

**Synchrony Financial (Fraud Analytics**) **Remote**

**Roles and Responsibilities:**

* The project is to identify the suspected fraud accounts in order to avoid financial losses.
* Imported data from data sources and performed spark transformations and actions on the data and stored the result in data lake.
* Imported data from multiple data sources using airflow scheduler DAG for one time and incremental.
* Created the Databricks pipeline for the extraction, transformation of data from different sources with scalability.
* Created Databricks notebooks using Scala/spark for transforming the data that is stored in Azure Data lake store and placed the final transformed results into the Azure blob.
* Did performance tuning for large datasets to make the code efficient.
* Extract, clean and process the data to normalize the feature to perform feature engineering on big data platform using spark and hive.
* Apply patterns code to identify different types of patterns for an account and generate fraud signals based on the selected identifiers by applying machine learning algorithms.
* Implemented different types of clustering on the accounts data to Identify the fraud rings and all the related accounts.
* Share the data insights with the business stake holders.
* Analyzed anomalies in Time Series data using various machine learning and deep learning models.
* Used python, Tableau to create visualizations for effective communication.
* Implemented PySpark jobs to apply business rules on the data to find the fraud accounts, later migrated these jobs to azure data bricks.
* Worked on Azure Data bricks to run Spark-Python Notebooks through ADF pipelines.
* Lead the effort to create design document and solution as per the business requirement

**Environment:** Tools: Azure Data Factory, Airflow, Azure Databricks, Scala, Spark, Azure SQL DB, Azure DW (Synapse Analytics), Azure DevOps, Storage account, Python, PySpark

**SR. DATA ENGINEER 03/2019 to 09/2019**

**Optum Technologies(HEDIS Quality measure engine) [Eden Prairie, Minnesota](https://en.wikipedia.org/wiki/Eden_Prairie%2C_Minnesota)**

 **Roles and Responsibilities:**

* This project was to build the Certified measures for HEDIS (Healthcare Effectiveness Data and Information Set) in order to do the star performance measure check of health plans for the member
* Development of HEDIS measures for members based on the facts and measure criteria given in an excel format using python, Java and spark
* Created Databricks notebooks using Scala for transforming the data that is stored in Azure Data lake store and placed the final transformed results into the Azure SQL.
* Created Big Query 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, Dataproc with Cloudera Distribution.
* Exposure to Azure Data Factory activities such as Lookups, Stored procedures, if conditions, for each, Set Variable, Append Variable, Get Metadata, Filter etc.
* Created automated pipelines in Azure Code Pipeline to deploy Docker containers in Azure ECS using services like CloudFormation, Code Build. Code Deploy, S3 and puppet.
* Created a Power BI data model based on analysis of the end-user workflow data provided by the client.
* Configured and implemented the Azure Data Factory Triggers and scheduled the Pipelines; monitored the scheduled Azure Data Factory pipelines and configured the alerts to be notified of failure pipelines.
* Created Big Query authorized views for row level security or exposing the data to other teams.
* Worked on Azure Databricks to run Spark-Python Notebooks through ADF pipelines.
* Create and architectural design for the audit output
* Designing and implementing product features in collaboration with business and IT stakeholders and working closely with the architecture group for driving technical solutions.

**Environment:** Tools: Azure Data Factory, Airflow, Azure Databricks, PLSQL, Kubernetes, Big Data, Scala, Azure SQL DB, Azure DW (Synapse Analytics), Polybase, Azure DevOps, Blob, ADLS, Power BI.

**SR. BIG DATA ENGINEER 11/2018 to 03/2019**

**Optum Technologies (Net Promoter Score)** Minnetonka, Minnesota

**Roles and Responsibilities:**

* This project was to improve the user experience for the services provided by determining NPS
* Provided end to end design, implementation and making project production ready
* Developed backend coding using Python and Azure server less to convert the XML data coming from S3 bucket into JSON Format and automated workflow of raising tasks and inserting the required data into DynamoDB and used previous data from DynamoDB and used to check if the errors are happening constantly.
* Contributed to the optimization and tuning of data processing pipelines and deployed ML models, improving performance and scalability.
* Designed and implemented scalable data processing pipelines using Python, focusing on libraries such as Pandas, Numpy, and Scikit Learn for data manipulation and analysis.
* Get the survey results from various sources and formats into the Hadoop environment
* Validate and clean the data and apply transformation rules.
* Transform and join the Member data with their respective survey data and saved it as a partitioned parquet file in Hadoop eco system
* Implemented validation framework using Spark SQL as part of data validation.

**Environments**: UNIX, CONTROL-M, QlikView, Azure, HIVE, Python, HDFS, DATASTAGE 11.3 and TERADATA

**SR. BIG DATA ENGINEER** *01/2018 to 09/2018*

## Optum Technologies (Centre IHR)

**Roles and Responsibilities:**

* Involved in the data flow and data analysis where this involves getting data from upstream and find the risk and quality measures for a given member
* Imported data from upstream into HDFS, HBase using Spark with Oozie workflows
* Designed and implemented framework to filter members using soft and hard logic
* Involve in gathering the individual's demographic, contact and Preference communications details and data profiling
* Design and Implemented Spark framework to process the HL7 and various other file formats and then push these to RabbitMQ
* Created ETL packages with different data sources (SQL Server, Oracle, Flat files, Excel, DB2, and Teradata) and loaded the data into target tables by performing different kinds of transformations
* Worked on importing and exporting data from Oracle, and DB2 into HDFS and HIVE using Sqoop for analysis, visualization, and generating reports.
* Worked on analyzing Hadoop clusters using different big data analytic tools including HBase database and Sqoop.
* Extract the data from HBase using spark Java/Scala API's and index into elastic search to present this in Kibana dashboards.

**Environment:** Hadoop, MapReduce, HDFS, Hive, python, Kafka, HBase, Sqoop, No SQL, Spark 1.9,

PL/SQL, Oracle, Cassandra, Mongo DB, ETL, MySQL

**SR. DATA ENGINEER 01/2016 to 01/2017**

## Optum Technologies (Consumer 360) Minnetonka, Minnesota

**Roles and Responsibilities:**

* This product was for making Active insurer's data easily available to the business people, Agents, advocates through Rest API's to view and update Individuals and Membership details
* Involve in gathering the individual's demographic, contact and Preference communications details and data profiling
* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating ETL pipelines using Sqoop, Hive and Spark.
* Responsible for importing data to HDFS using Sqoop from different RDBMS servers and exporting data using Sqoop to the RDBMS servers after aggregations for other ETL operations.
* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating ETL pipelines using Sqoop, Hive and Spark.
* Experience in designing and developing applications in PySpark using python to compare the performance of Spark with Hive.
* Generate Java Components from XSD and extract the data from HBase
* using spark java and Scala API's and Hive SQLs, map the extracted data to Java components generated
* Experience in designing and developing applications in PySpark using python to compare the performance of Spark with Hive. Lead the end-to end effort at Hadoop environment
* Maintaining users, roles, and permissions within the SQL Server environment, enforcing proper data access controls and security measures.

Environment: SQL Server 2008/2012 Enterprise Edition, SSRS, SSIS, T-SQL, Windows Server 2003, Performance Point Server 2007, Oracle 10g, visual Studio 2010.

**Environment:** Hadoop, Spark, Hive, Impala, HBase, ETL, Sqoop, Hive, Spark, Azure SDKs, Python, PySpark, Datadog API, Docker, Kubernetes (EKS), EMR, EC2, S3, DynamoDB, Lambda

**HADOOP DEVELOPER 08/2015 to 12/2015**

## Target (Price Publish) Minneapolis, Minnesota

**Roles and Responsibilities:**

* This project gathered various optimized prices for an item from different systems
* Apply analytics to determine the current price required for an item
* Involved in design and implementing the routes, processors, aggregators for apache camel
* Implemented application with spring boot along with elk setup with chef on
* open stack cluster
* Coordinated with Hadoop admin on cluster job performance and security issues, and Hortonworks team to resolve the compatibility and version related issues of HDP, Hive, Spark, Oozie.
* Built the summary tables, implemented call prediction models: player gaming summary models with K-Means Cluster in production using Spark MLlib and Scala.
* Worked extensively with Sqoop for importing and exporting data from SQL Server.
* Implemented Preprocessing steps using Data Frames for Batch Processing
* Worked with Data scientist partner for Predictive analysis, implemented bonus recommendation Engine using Spark MLib, and persisted the recommendation results in HBase
* Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into HBase
* Represent the data on Kibana dashboard that shows system data and metrics.

**Environment**: Hadoop, MapReduce, HDFS, Hive, python, Kafka, HBase, Sqoop, No SQL, Spark 1.9,

PL/SQL, Oracle, Cassandra, Mongo DB, ETL, MySQL

**HADOOP DEVELOPER 02/2015 to 07/2015**

## Target (Context Determination Engine) Minneapolis, Minnesota

**Roles and Responsibilities:**

* Determining the optimized price for an item by taking competitors data and applying rules to it
* Price lookup for the threshold of an item price and notify the pricing applications
* We calculate the goal retail for an item and check if that item is in clearance or sale or any promotion plans then apply the promotion or sale price till the specified last date, if not then apply regular retail price
* Creating and inserting data into Hive tables for dynamically inserting data into data tables using partitioning and bucketing for EDW tables and historical metrics.
* Created ETL packages with different data sources (SQL Server, Oracle, Flat files, Excel, DB2, and Teradata) and loaded the data into target tables by performing different kinds of transformations
* Installed and configured Hadoop MapReduce HDFS Developed multiple MapReduce jobs in java for data cleaning and pre-processing.
* Developing Spark scripts, UDFS using Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.
* Spark, Effective &amp; efficient Joins, Transformations, and others during the ingestion process itself.
* Apply business rules using drools in the storm bolt Publish the data on a kafka queue
* Analyze the input data coming from various sources and designed the schema to ingest the data.

**Environment:** Hadoop, MapReduce, HDFS, Hive, python, Kafka, HBase, Sqoop, No SQL, Spark 1.9, PL/SQL, Oracle, Cassandra, Mongo DB, ETL, MySQL

**HADOOP DEVELOPER** *03/2013 to 12/2014*

## Adaptive (Real World Analytics) - Hyderabad, India

**Roles and Responsibilities:**

* This product focus on data driven marketing that helps leading retail organizations grow their businesses
* Implemented Spark SQL queries that combine hive queries with Python programmatic data manipulations supported by RDDs and data frames.
* Developing Spark scripts, UDFS using Spark SQL query for data aggregation, querying, and writing
	+ data back into RDBMS through Sqoop.
* Installed and configured Hadoop MapReduce HDFS Developed multiple MapReduce jobs in java for
	+ data cleaning and pre-processing.
* Installed and configured Pig and also written Pig Latin scripts.
* Gets a precise measure buying behavior at the individual customer level and provides a detailed analysis of how marketing impacts revenue
* Designed and implemented scalable data processing pipelines using Python, focusing on libraries such as Pandas, Numpy, and Scikit Learn for data manipulation and analysis.
* Provide trends for different products with regard to customer behavior
* Worked with data scientists to provide data as per their requirements
* Utilized SQL for data querying and manipulation, extracting and transforming data from various sources for ML model training and evaluation.
* Worked on Azure glue ETL tool, in which we used Azure Glue Crawler, Azure Glue Data catalog and
	+ Connections.
* Designed and implemented to ingest and store Meta data from various data providers
* Worked on designing and developing the SSIS Packages to import and export data from MS Excel, SQL Server, and Flat files.
* Analyzed the data by performing Hive queries and running Pig scripts to know customer purchase behavior
* Utilized Oozie workflow engine to run various Sqoop, Hive and Pig jobs.

Environment: Hadoop, MapReduce, HDFS, Hive, python, Kafka, HBase, Sqoop, No SQL, Spark 1.9,

PL/SQL, Oracle, Cassandra, Mongo DB, ETL, MySQL

**JAVA DEVELOPER** *08/2011 to 05/2012*

## HSBC, Mortgage Services - Hyderabad, India

**Roles and Responsibilities:**

* CML Data warehouse systems were developed to meet the business reporting and analytical requirements by the Marketing CIM Team
* Extensively used MVC, Factory, Delegate and Singleton design patterns.
* Involved in Requirement gathering, Analysis, Preparing Functional Specifications, and technical specifications.
* This application determines the delinquent customers based in foreclosure data who will be given a chance to avoid foreclosure by settling the loan with liquidating the physical assets or paying off the partial loan amount by submitting the details in application.
* After analysis and processing stage marts are created which are sent to the
* Customer Information Management and Reporting team for analysis purpose and making better decisions.

Environment: JDK1.5, JSP, Servlet, JSTL, JavaScript, Spring IOC, Spring MVC, Spring DAO, Hibernate,

Web Logic Server, Log4j, xml, xsl, Jaxb, Jaxp, Maven, Oracle, Toad, CVS, Eclipse, and Mercury.

**MAINFRAME DEVELOPER** *04/2010 to 08/2011*

## One HSBC, Credit and Retail Services - Hyderabad, India

**Roles and Responsibilities:**

* CRS DW provides business decision makers the ability to perform analysis as over a period of time on customer, product profitability, risk analysis, market trends through various data marts
* Developed GUI using HTML, JavaScript, CSS and JSP pages using standard and custom tags.
* Spring MVC used in the presentation layer for commercial Sales application.
* Struts and tiles used for privacy law compliance survey application.
* Different J2EE Design Patterns applied for projects.
* Used XML for transferring data between the servers using XML OVER HTTP.
* Gathered and Analyze Requirement from the business managers for change request (CR) and New Development.
* Provide Support to the Production Team.
* This solution helps HSBC Technological Services achieve tactical and strategic goals by enabling them with solutions which generate information & vital statistics to make the right decisions.

**Environment:** JDK1.5, JSP, Servlet, JSTL, JavaScript, Spring IOC, Spring MVC, Spring DAO, Hibernate,

Web Logic Server, Log4j, xml, xsl, Jaxb, Jaxp, Maven, Oracle, Toad, CVS, Eclipse, and Mercury.

**MAINFRAME DEVELOPER** *06/2007 to 03/2010*

**ASSURANT -** Hyderabad, India

**Roles and Responsibilities:**

* BEST (Better Examiner System Technology) is a medical and dental claim processing system, which captures claims, adjudicates and pays medical and dental insurances
* It receives claims from manual and online sources and attempts to adjudicate them without manual intervention, passing the approved claim onto back-end systems that pay the claim
* When the processor has finalized the claim, a record is added to the claim transaction file which is processed in the nightly batch cycle.