**Name**

**Azure Data Engineer**

Email:

Linkedin:

Work Authorization:

**SUMMARY**

* **5**+ Years of experience in **Microsoft Azure/Cloud Services** like **Synapse, Azure SQL Server, Azure Databricks, Azure Data Lake, Azure Blob Storage, Azure Data Factory, CosmosDB, EventHubs**
* Over **8+** years of experience in system analysis, design, development and testing in various projects,
* **2**+ years working with Big Data platforms like **Apache Spark, Hadoop, Hive**
* **2+ years of experience in streaming analytics Spark Streaming, Dstream, Databricks Delta**
* **2+ years of experience in Databricks MLflow for running machining learning models on distributed platforms**
* Expertise in programming tools like **Python, R, SAS** and complete maintenance in Software Development Life Cycle (SDLC) like **Agile/SCRUM environment**.
* Extensive experience in handling **Large Datasets of Structured and Unstructured data and building Models using various Machine Learning Algorithms and Techniques**.
* Adaptable in various Data Modeling packages like **NumPy, SciPy, Pandas, Beautiful Soup, Scikit-Learn, Matplotlib, Seaborn in Python and Dplyr, TidyR, ggplot2 in R**.
* **Deep understanding of various Machine Learning Algorithms** like Linear and Logistic Regression, Decision Trees, Ensemble Methods, Recommendation systems, Clustering, Time series and Forecasting, Predictive Modeling, **PCA** , **Neural Network** algorithms to train and test the huge data sets.
* Experience in working with **NoSQL Databases** like **CosmosDB**, **MongoDB**
* Over **2**+ years of experience in **IBM DataStage** design and development of parallel ETL jobs
* Over **2**+ years of experience in **Unix shell scripting**, for ETL job run automation and Datastage engine admin
* Implementation of **Master Data Management** in **Data Warehouse** ensuring data governance, removal of **duplicate records**, data cleansing and profiling.
* Experience writing deployment script using **DOCKER**, **SQLCMD**, **RS**, **DTUTIL** and release document for release management team.
* Worked extensively on system **analysis**, **design**, **development**, **testing** and **implementation** of projects (Complete **SDLC**) and capable of handling responsibilities independently as well as a proactive team member.
* Good SQL Server Administration skills including **backup recovery, database maintenance, user authorizations, Database creation, Tables, indexes, Partitions, running Database Consistency Checks using DBCC**.

**PROFESSIONAL SUMMARY**

**Premera BlueCross May 2020 – Dec 2020**

**REMOTE**

**Senior Data Warehouse Engineer**

Consulting as an Azure Data Engineer and architect for Premera Health Insurance Company. Utilizing Microsoft Azure tools to build out a Data warehousing with Snowflake Datawarehouse. Helping them migrate from on premise Netezza and Teradata data stores to the cloud.

**Responsibilities:**

* Create an Architectural Solutions for ingesting Data using DataStage to Azure Datalake Gen 2 Staging environment
* Setting up a **Snowflake** Datawarehouse Units for Business Intelligence reporting
* Build shared utility functions wrapped on azcopy for file transfers to cloud
* Remediating existing code base to switch from on Premise Netezza target database to Snowflake target
* Create visual basic scripts for ETL processing and pipelines
* Transformed and manipulated **unstructured data (json)** using pyspark, python pandas
* Create data validation and ETL validation checks to ensure proper data flow
* Create dimensional data model for analytics
* Design and present technical solutions to end users in a way the is easy to understand and buy into
* **Educating client/business users on** the pros and cons of various **Azure PaaS and SaaS** solutions ensuring the **most cost-effective** approaches are taken into consideration
* **Create PowerBI Data models and Power Query for optimal Report performance**
* Create **Self Service** reporting **in Azure Data Lake Store Gen2** using an ELT approach.
* Designing ELT pipelines for Facts and Dimension table loading into the **Snowflake** Datawarehouse
* Implementing Real Time data ingestion from Source Databases using **Attunity** CDC Replication tool
* Setting up Azure infrastructure like **storage accounts, integration runtime, service principal** id, app registrations to enable **scalable and optimized** utilization of business user analytical requirements in Azure.
* Writing SQL queries to transform data and load them into **Snowflake**
* Creating Batch jobs for loading data into the data warehouse in Netezza and Teradata
* Migrating Data from Netezza to Azure Snowflake Warehouse
* Deploying code via Azure Devops as CI/CD methodology

**Tools:** Azure Data Lake Gen2, Datastage, SnowFlake, Azure Data Factory, Spark, Databricks, Azure Devops, Agile, PowerBI, Python, R, SQL, Scaled Agile team environment, Autosys

**Publicis Sapient Sep 2019 – April 2020**

**Houston, TX**

**Senior Data Engineer**

Consulting as an Azure Data Engineer and architect for Chevron Health Environment Safety Business unit. Utilizing Microsoft Azure tools to deliver analytics that support key business users in making decisions

**Responsibilities:**

* Create and maintain optimal data pipeline architecture in cloud **Microsoft Azure using Data Factory and Azure Databricks**
* Create an **Architectural solutions** that leverages the best Azure analytics tools to solve our specific need in Chevron use case
* Design and present technical solutions to end users in a way the is easy to understand and buy into
* **Educating client/business users on** the pros and cons of various **Azure PaaS and SaaS** solutions ensuring the **most cost-effective** approaches are taken into consideration
* **Create PowerBI Data models and Power Query for optimal Report performance**
* Create **Self Service** reporting **in Azure Data Lake Store Gen2** using an ELT approach.
* Utilize **Azure Key Vault** to secure and protect authentications against other Azure services and store keys
* Create **Spark Vectorized panda user defined** functions for data manipulation and wrangling
* Transfer data in logical stages from System of records to raw zone, refined zone and produce zone for easy translation and denormalization
* Setting up Azure infrastructure like **storage accounts, integration runtime, service principal** id, app registrations to enable **scalable and optimized** utilization of business user analytical requirements in Azure.
* Writing pyspark and spark sql transformation in Azure Databricks to perform complex transformations for business rule implementation
* Creating Datafactory pipelines that can bulk copy multiple tables at once from relational database to Azure datalake gen2
* Create custom logging framework for ELT pipeline logging using Append variables in Datafactory
* Enabling monitoring and azure log analytics to alert support team on usage and stats of the daily runs
* Took proof of concept projects ideas from business, lead, developed and created production pipelines that deliver business value **using Azure Data Factory**
* Kept our data separated and secure across national boundaries through multiple data centers and regions.
* Implement Continuous integration/continuous development best practice using Azure Devops, ensuring code versioning
* Utilized **Ansible playbook** for code pipeline deployment
* Delivered denormalized data for **PowerBI** consumers for modeling and visualization from the produced layer in Data lake
* Worked in a **SAFE (Scaled Agile Framework)** team with daily standups, sprint planning, quarterly planning.

**Tools:** Azure Data Lake Gen2, Azure Data Factory, Spark, Databricks, Azure Devops, Agile, PowerBI, Python, R, SQL, Scaled Agile team environment

**Henkel Aug 2018 – Aug 2019**

**Rocky Hill, CT**

**Data Scientist/Senior Data Engineer**

Henkel is a leading global chemical manufacturing company with various product lines from Adhesives to beauty products.

**Responsibilities:**

* Create and maintain optimal data pipeline architecture in cloud **Microsoft Azure**
* Utilized web scraping techniques like **Beautiful Soup** in Python to extract and organize competitor data
* Performed **Time series analysis** and **decomposition** (**ARIMA, STL, ETS)** on product sales records to understand seasonality and trend of in store Hair product sales
* Developed a prediction model using **Xgboost and Random Forest** to predict the number of units sold given independent variables like week, unit price, store count
* Recommended Sales Promotional strategy based on in depth analysis of sales trend and model predictions
* Migrated projects from Cloudera Hadoop Hive storage to Azure Data Lake Store to satisfy Henkel Digital transformation strategy
* Took proof of concept projects ideas from business, lead, developed and created production pipelines that deliver business value **using Azure Data Factory**
* **Implement IOT streaming with Databricks Delta tables and Delta lake to enable ACID transaction logging**
* Exposed transformed data **in Azure Spark Databricks platform** to parquet formats for efficient data storage
* Architected solutions for various analytics requirements in the cloud that were scalable and efficient utilizing SQL Databases and **ELT** techniques
* Assemble large, complex data sets that meet functional / non-functional business requirements.
* Implemented **Spark SQL** jobs for data transformation
* Built the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using SQL and ‘big data’ technologies like **Hadoop Hive, Azure Data Lake storage**
* Built the data platform for analytics tools that utilize the data pipeline to provide actionable insights into customer acquisition, operational efficiency and other key business performance metrics.
* Work with stakeholders including the Executive, Product, Data and Design teams to assist with data-related technical issues and support their data infrastructure needs.
* Kept our data separated and secure across national boundaries through multiple data centers and regions.
* Implement **Tableau server user access control** for various Dashboard requirements
* Implement complex business rules in **python using Panda libraries, numpy, Scikit learn**
* Optimize **ELT work loads against Hadoop file system** implementing HIVE SQL for transformation
* Containerize data wrangling jobs in **Docker containers** **utilizing Git and Azure DevOps** for version control

**Tools:** Hadoop, Hive, Azure Data Lake, Azure Data Factory, Spark, Databricks, Dremio, Tableau, PowerBI, Python, R, Knime, Docker

**Northwell Health System Aug 2016 – Aug 2018**

**New Hyde Park, NY**

**Data Scientist/Senior Data Engineer**

Northwell Health System is one of the largest integrated health systems in the United States; specializing in in-patient and out-patient care. The main aim of the project was to implement IBM Unified Health Care Data model, to build a data warehouse with the help of an accelerator

**Responsibilities:**

* Worked with Perficient Consulting group to build an Enterprise Data warehouse solution for Northwell Health
* Developed **predictive models** with Clinical Analysts for LACE score index calculations with **R machine learning algorithm**
* **Performing data transformation and wrangling using python scripts**
* Performed Statistical analysis using **linear regression** to find correlation between data points
* **Inferred model results based on multivariate prediction models like K nearest neighbors, logistic regression, Naïve Bayes etc. with Bayesian inference method.**
* Analyzed and processed complex datasets about **patient readmission rates** and ways to reduce it.
* Designed and developed DataStage code for extracting source data into Persistent stage layer then to a Standard Interface gateway **using SCDII (type 2 slowly change dimension) methodology** and Microsoft Change Data capture
* Developed Spark programs using **Python (PYSpark)** API for data wrangling and transformations.
* Queried Structured Data using **Spark SQL** in order to enable rigorous analysis
* In depth understanding/knowledge of **Hadoop** Architecture and various components such as HDFS, Application Master, Node Manager, Name Node, Data node and MapReduce concepts
* Enhanced the vendors ETL framework to ensure a more dynamic and parameterized design to extract data from source tables using ETL configuration tables
* Designed about **ELT** jobs to move, integrate and transform **Big Data** from various sources to a single Target database.
* Scraped webpages using XPATHs, API querying, using parsing frameworks like BeautifulSoup, and LXML
* Used **Regular expression (Regex)** in **Python and SQL to extract** and extrapolate vital medical metrics for analytics from lab results notes.
* Maintained over **100 ETL jobs** that populate IBM UDMH (Unified Data Model for Healthcare) **Atomic Model** and **Dimensional model**
* Extensive understanding of  **Linstedt** **Data Vault** modeling and ETL loading implementation and architecture
* Designed ETL’s to populate **Anchor tables**, **array** tables, **bridge** tables, **Detail** tables and **Reference** tables
* Loaded data into **Netezza** database, ensuring proper distribution of data across data slices with the most adequate **distribution keys**
* Converted data mapping documents to ETL/Datastage jobs loading data into Netezza.
* Created **Unix scripts** to execute jobs automatically and in a schedule
* **Debug** and monitor parallel job failures and errors that arise from source system anomalies or interepratation of data mapping
* Designed ELT jobs to load Big Data over 4 billion records from SQL Server Data Source to Massively Parallel Processing (MPP) IBM Pure Data Appliance
* Performed various performance tuning techniques to leverage the columnar data storage, and distributed computing architecture of Netezza DWH appliance
* Implement **CDC from Source system** to Atomic and Dimensional tables, ensuring hard and soft deletes are flagged at the record level
* Create **Data Validation** scripts that would be used for verification correctness of ETL logic and transformations between source system tables and dimension tables.

**Northwell Health System Aug 2015 – Aug 2016**

**New Hyde Park, NY**

**Data Engineer**

The main aim of the project was to design, and propagate changes from SQL Server database source using Microsoft Change data capture and IBM Datastage to a STAGE Oracle database. This STAGE database will be subsequently used as a source connection for a Health system BIG Data analytics vendor.

**Responsibilities:**

* Designed and developed a **SCD I incremental loading** process from a SQL Server Data Source to a Oracle Staging Database using **IBM Infosphere Change Data Capture**, PLSQL packages and **Datastage**
* **Set up Infosphere Change Management console subscriptions** to a SQL Server database,set bookmarks to start of the **Log Reading and configured** Oracle Target Tables for record insert
* Designed and Developed **DataStage jobs** to process **FULL Data loads** from SQL Server Source to Oracle Stage
* **Imported Table Definitions** using **ODBC** plug in in **DataStage Repository**
* Designed one **Parameterized** job using **Parametersets** that can be re used for multiple tables implementing **Run Time Column Propagation** in **DataStage**
* Created **Datastage** jobs that wrote into **Parameter files** so that subsequent jobs in the **sequence** can read it for proper execution
* Designed **Data Stage SEQUENCE Jobs** that controlled ETL for multiple tables in a subject area and sends a success email once job is complete
* Implemented complex **Datastage Transformer logic** for various business rules in the ETL
* Designed and developed incremental load logic that used a **control table that store the min and max LSN(transaction commit cycle ID)** for the successfully loaded transactions.
* Implemented **Error Handling in Datastage** and designed Error jobs to notify user and update log table
* Designed performance boosting jobs that ran the Datastage job on 4 nodes taking **advantage of Datastage Parallel execution of different** partitions to reduce job run time
* Implemented **Microsoft SQL Server Change Data Capture for** SQL Server Data sources, taking advantage of the inbuilt functions like sys.get\_min\_LSN, sys.get\_max\_LSN, and sys.net\_change
* Installed SQL Server 2012 and Management tools using SQL Server Setup Program.
* Created **Unique index on business keys** in table enabling data validation and integrity.
* Designed and Developed an **ETL** **Control Log table** that records the number of inserts, update ,delete and error message for each running process
* Created Tables in Oracle specifying the appropriate **tablespace**, **data storage parameters( initial extent, next extent, pct increase)**

**Northwell Health System Aug 2014 – Aug 2015**

**New Hyde Park, NY**

**Senior Data Warehouse / IBM DataStage Developer**

The main aim of the project was to design, develop and populate a STAGE Oracle database for data integration for 12 different sources with historical data.

* Understood **business/data transformation rules,** business structure and hierarchy, relationships, data transformation through mapping development.
* Pulling Data from **MDM** data repository to Data warehouse engine
* Reverse Engineered the Stage Schema to create a Data Model using **Embarcadero ER/Studio Architect.**
* Created a **Data Lineage** from Source Table to Stage Tables showing data mapping rules, transformations and business logic using Embarcadero ER/Studio Architect.
* Developed **SSIS packages** to extract data from SQL Server Source and load into Oracle Database using **Oracle Attunity Drivers** Develop **PL/SQL packages**, procedures, functions, and views for business requirement
* Worked extensively on **Ref Cursor, External Tables and Collections**.
* Implemented large table partitions using **RANGE Partition and LIST Partition techniques**
* Implemented and design procedures using **FOR ALL and BULK Collection concepts**
* Experience with **Performance Tuning** for Oracle RDBMS using **Explain Plan** and **HINTS.**
* Expertise in **Dynamic SQL, Collections**and**Exception handling**
* **Implemented Partition Exchange to migrate data for over 2 billion records efficiently from MS SQL Server to Oracle DB**
* **Created Stage tables to process Source data per logical partition for ETL Parallelism in SSIS.**
* **Created an SSIS Package to efficiently load CLOB columns from SQL Server to Oracle using Attunity drivers and Conditional Split transformation**
* Created Source to Target Mapping Documentation in excel to outlines the data conversion/transformation routines implemented from SQL Server Source tables columns to Oracle target Table columns
* Created Source analysis documentation to describe the tables in source system with relationships outlining business keys
* Modified table parameters to enable faster bulk load with **NOLOGGING and Parallel hint**
* Enable parallel DML insert (**Direct Load INSERT**)
* Developed SSIS packages to extract data from MS SQL DB to Oracle DB for over **2 billion** records and **1 TB** data volume
* Developed **Oracle Packages** with Procedures for Full and Incremental ETL processes using Merge statement
* Implemented **SCD II** in **incremental** loading, keeping the historical track of Inserted, updated and deleted records using **SQL MERGE STATEMENTS,** ensured to process only changed records
* **Implemented SCD I in incremental** loading of huge historical table from source to a stage environment
* Modified and validated **Audit Trail Triggers capturing CDC** records.
* **Partitioned tables** with size over 100GB to increase availability and efficiency of full and incremental load into stage environment.
* **Created Mapping documents for data analytics vendor for various Source systems**
* **Created Source to Target and Source Analysis documentation for various Source Systems imported to Stage Environment**
* Implemented **Parallel data** r**eading** with separate data flows running in parallel and loading into one table
* Participated in **daily Stand up meeting** working on **Product Back log for Sprint goals**

Engaged in 3 week sprint in **Scrum Development methodology** in fast paced work envir

**Environment:** MS SQL SERVER 2014/2012, **SSIS**, SSRS, Business Intelligence Studio ,DTS, Oracle12C, T-SQL, VSS 2005, Erwin r7.2, SQL profiler, Embarcadero ER Studio Data Architect, Dot NET framework 3.5, ASP.NET,C#, XML, PL/SQL, Star and Snowflake schema Data Models

**EDUCATIONAL QUALIFICATION**

Bachelors of Electrical Engineering City College of New York, NY   **GPA 3.88**