**Nitisha Chilaka ** 

**Email: nitisharajchilaka@gmail.com** **| Ph: 609-297-5221 |Location: Jersey City, New jersey |**

 **SENIOR DATA ENGINEERING PROFESSIONAL ~ 11+ Years of Experience**

**Big Data | Data Science | Cloud Engineering | Machine Learning | Automation**

 **Profile Synopsis**

* + Verifiable experience in leveraging Data Lake, Cloud platforms & BI layers
	+ Excellent track record of managing high level engagements within AI, ML, and Data Science domains across geographies
	+ Solid exposure to leading legacy application revamps based on thorough assessments of new requirements
	+ Verifiable experience in ETL data processing through Hadoop Native tools / NIFI / Informatica BDM.
	+ Solid exposure to Real time data processing with Kafka & MQTT.
	+ Handled many data migrations from multiple source systems (IBM mainframe, MYSQL, Oracle) to Hadoop through sqoop.
	+ Knowledge and working experience on big data tools like Hadoop, Azure Data Lake, AWS Redshift. Good understanding of Apache Airﬂow.
	+ Solid experience in Data orchestration tools like Apache Oozie.
	+ Automated many process / data pipelines through scripting, Jenkins
	+ Supported Modelop image deployments / Creating image for data science teams based on the requirements.
	+ Strong domain knowledge in Retail, Banking, IOT, Product, Finance, and CRM Platforms.
	+ Collaborative team player with excellent interpersonal, communication, organizational, and leadership skills.
	+ Deployed scalable, containerized data pipelines on OpenShift, optimizing resource usage and enabling seamless scaling for high-volume data processing tasks.
	+ Experienced in conﬁguring and administering the Hadoop Cluster using major Hadoop Distributions like Apache Hadoop and Cloudera.
	+ Designed and implemented scalable cloud architectures using Google Cloud Platform (GCP) services, optimizing performance, cost, and security for enterprise-level applications.
	+ Deployed and managed high-availability applications using GCP Compute Engine and App Engine, leveraging autoscaling features for efficient resource utilization.

#  Skills

* **Techniques:** Data Analysis, Data modeling, Data Mining, Statistics, Data Visualization, Data Management, Clustering, PCA, Regression, Decision Tree, Random Forest, Forecasting, Supply Chain, Qualtrics.
* **Programming Languages:** Python, SQL, R-Programing.
* **Data Visualization Tools**: Tableau, Power BI, Qlik, JMP, Salesforce CRM
* **Product Management Tools**: Jira, Azure DevOps, BIA, BCP.
* **Database Management**: MySQL, Oracle12c, PostgreSQL
* **Tools**: MySQL workbench, SQL.
* **Operating System**: Windows (8, 10,11), Ubuntu, Linux
* **SDLC:** Agile, Waterfall
* **Reporting/Data modeling:** Tableau, Cognos 8 Suite, Star Schema, Snowflake- Schema, Fact and dimension tables, Pivot tables, Erwin, CRM
* **Defect Tracker**: Quality Center, HPQC, HP-ALM, SF-toolkits JIRA, ServiceNow
* **Languages:** Python, SQL
* **Big Data/Hadoop technologies:** Snowflake, Apache Hadoop (MR1, MR2), Apache Spark, Kafka, Apache NiFi, Flume, Knox, Ranger, Atlas, Hive, Impala, Hbase, Impala-Kudu, Sqoop, Oozie workflow, HDFS, Yarn, Informatica BDM
* **Cloud Services: GCP, AWS Cloud Tools:** Amazon EC2, Amazon EMR, ELB, Amazon Lambda, Amazon Redshift & S3.
* **Azure Services**: Azure Data Factory, Azure SQL, Azure Datalake
* **GCP Services:** Cloud SQL, Cloud Functions, GCP Cloud Storage, Big Query
* **Message Broker Tools:** Kafka, MQTT
* **Operating Systems:** Mac, Linux & Windows | Databases: NoSQL (HBase), SQL (Hive, Impala, MySQL), Presto DB
* **CI/ CD tools:** Jenkins, Docker, Kubernetes
* **Cloud Platforms**: OpenShift, AWS, GCP, Azure

**Professional Experience**

**Client: Fidelity Investments, Jersey City, NJ OCT 2023 – PRESENT**

**Title: Senior Data Engineer/Technical Lead**

**Responsibilities:**

* + Process management - Ensuring adherence to the organization & development process as per the WBS (work-based structure), maintaining PMT as per the estimations made.
	+ Assisted Project Manager to create Project Charter, Project Plan Document, Project Road Map, Resource Allocation etc.
	+ Developed a real-time generic alert structure process for easy analysis with Kafka, Python & Snowflake.
	+ Automated the Rehydration process for docker images, Modelops images through Jenkins and snow API.
	+ Design and Develop ETL Processes in AWS Glue to migrate data from external sources like S3, Parquet/Text Files into AWS Redshift.
	+ Developed and Maintained SharePoint Sites: Designed and implemented custom SharePoint solutions to support data engineering and business intelligence workflows.
	+ Experience building Data pipeline for Realtime streaming data and Data Analytics using Azure cloud components like Azure Data Factory, HDInsight (spark cluster), Azure ML Studio, Azure stream Analytics, Azure Blob Storage, Microsoft SQL DB, Neo4j (Graph DB).
	+ Implemented Apache Iceberg tables for scalable, high-performance, and versioned data sharing across data lakes and analytics platforms.
	+ Designed and implemented data solutions using Microsoft Fabric, enabling seamless data integration and analytics across the organization.
	+ Conducted data cleaning, transformation, and analysis using Pandas, PySpark, and Databricks notebooks.
	+ Experience in building ETL(Azure Data Bricks) data pipelines leveraging PySpark, Spark SQL
	+ Managed MongoDB clusters for high availability and disaster recovery.
	+ Designed and implemented scalable data pipelines using AWS Glue and AWS Lambda to automate ETL processes.
	+ Designed and implemented robust data pipelines using DBT, improving data processing efficiency and reliability.
	+ Experienced in developing end to end automation using Selenium WebDriver/RC/IDE/Grid, Unittest/Pytest, Jenkins, GHERKIN/Cucumber, Robot, ALLURE reporting, RESTful API and PostMan.
	+ Proficient in designing and implementing data integration pipelines using StreamSets Data Collector to efficiently manage and process data flows.
	+ Built scalable ETL pipelines in Microsoft Fabric using Dataflows and Pipelines to ingest, transform, and share data across multiple environments.
	+ Developed and maintained Iceberg table schemas to ensure compatibility with evolving data models and downstream applications.
	+ Built and deployed a containerized ETL pipeline using OpenShift to process large datasets from multiple sources in real-time.
	+ Assisted in building real-time data processing solutions using Apache Flink and Apache Kafka, gaining hands-on experience in stream processing and event-driven architecture.
	+ Managed and optimized big data processing workflows using Cloud Dataproc, integrating Hadoop and Spark clusters for large-scale data analytics.
	+ Automated Data Workflows: Created SharePoint workflows and integrated Power Automate for data extraction, processing, and task automation.
	+ Created and implemented ETL pipelines using Apache NiFi and Talend, facilitating seamless data flow from diverse sources into Hadoop.
	+ Developed and optimized data lakehouse solutions within Microsoft Fabric's OneLake architecture for unified storage and collaboration.
	+ Designing Data Models for Data Warehouses and Data Lakes, Optimizing Data for Performance and Storage
	+ Proficient in writing and debugging complex Linux shell scripts for task automation.
	+ Handled Hadoop cluster installations in Windows environment.
	+ Collaborated with data scientists to integrate machine learning workflows into Dagster pipelines.
	+ Designed and implemented data workflows using Prefect, ensuring robust and reliable data pipelines.
	+ Conducted data cleaning, transformation, and analysis using Pandas, PySpark, and Databricks notebooks.
	+ Experience in building ETL(Azure Data Bricks) data pipelines leveraging PySpark, Spark SQL
	+ Integrated Druid with Hive for High availability and provide data for sla reporting on real time data.
	+ Extensive experience in developing Kafka producers and Kafka consumers for streaming millions of events per minute on streaming data using PySpark, Python & Spark Streaming.
	+ Data Integration: Connected SharePoint with external data sources like SQL Server, APIs, and cloud platforms to enable seamless data flow.
	+ Proficient in Java programming for data processing and ETL pipeline development.
	+ Automated the provisioning of GCP infrastructure using Terraform, enabling consistent and repeatable deployment of cloud resources with infrastructure as code (IaC) best practices.
	+ Led data migration efforts from on-premise systems to GCP, utilizing Google Cloud Storage Transfer Service and Data Transfer Appliance for secure and efficient data transfers.
	+ Integrated Scala-based applications with Apache Spark to perform large-scale data processing and transformations.
	+ Proficient in utilizing data structures like arrays, linked lists, stacks, queues, trees, and graphs to optimize data storage and retrieval.
	+ Ensuring data transformation processes align with the data model, preserving relationships, and maintaining data integrity.
	+ Designed and implemented real-time data processing pipelines using Apache Flink, achieving sub-second latency for data ingestion and analysis.
	+ Extensive experience in developing Kafka producers and Kafka consumers for streaming millions of events per minute on streaming data using PySpark, Python & Spark Streaming.
	+ Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL
	+ Coordinated with multiple stakeholders to ensure that test data needs are addressed proactively, appropriately testing critical business systems with production like data which does not comprise of any PHI/PII.
	+ Curate data sourced out of Lake in to Databricks in different phases environments and perform Delta strong data engineering experience in Spark and Azure Databricks, running notebooks using ADF.
	+ Managed data storage and processing in cloud environments (AWS and GCP) using OpenShift, ensuring high availability and fault tolerance.
	+ 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.
	+ Implemented hybrid cloud solutions using Cloud Interconnect and Cloud VPN, enabling secure, low-latency connectivity between on-premises data centers and GCP.
	+ Wrote and optimized complex SQL queries to retrieve and manipulate data from relational databases.
	+ Implemented Oracle PL/SQL packages, triggers, and views to support business requirements.
	+ Leveraged Prefect's dynamic task mapping to handle complex data dependencies and workflows.
	+ Implemented real-time data streaming and processing solutions using Amazon Kinesis and AWS Lambda.
	+ Developed and maintained documentation for data architecture, including data flow diagrams and metadata repositories.
	+ Optimized query performance on Iceberg tables by leveraging advanced partitioning, compaction, and indexing strategies.
	+ Developed APIs and data sharing solutions to enable seamless integration between data systems, ensuring high availability and reliability.
	+ Integrated Microsoft Fabric with Azure Data Services (e.g., Azure Synapse, Azure Data Factory) to enhance data processing and analytics capabilities.
	+ Designed, developed and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems.
	+ Expertise in Core Java, data structures, algorithms, Object Oriented Design (OOD) and Java concepts such as OOP Concepts, Collections Framework, Exception Handling, I/O System and Multi-Threading.
	+ Design, develop, deploy and maintain large scale Tableau dashboards for Product Insight, Devices and Networking, and Cox Premise Equipment.
	+ Used PySpark jobs to run on Kubernetes Cluster for faster data processing
	+ Developed the ETL module / Front end module for Investigation workbench project.
	+ Supported production deployments, Docker image promotions.
	+ Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and Controlling and granting database access and Migrating On premise databases to Azure Data lake store using Azure Data factory.
	+ Designed and deployed secure, scalable APIs using GCP API Gateway, ensuring seamless integration with backend services while enforcing strict access control policies.
	+ Wrote and optimized advanced SQL queries, including joins, subqueries, and window functions for PostgreSQL.
	+ Customized Dagster configurations to optimize resource usage and processing speed for large datasets.
	+ Designed, developed, and maintained ETL processes using Informatica PowerCenter to integrate data from various source systems into data warehouses.
	+ Implemented the Big Data solution using Hadoop, hive and Informatica to pull/load the data into the HDFS system.
	+ Using tools like Erwin, Lucidchart, or dbt for designing and managing data models visually and ensuring all dependencies and constraints are documented.
	+ Configured and managed Talend projects, repositories, and job execution plans.
	+ Collaborated with cross-functional teams to integrate Databricks workflows with other AWS services for seamless data processing.
	+ Utilized Microsoft Fabric's Data Engineering features, such as Spark-based notebooks and pipelines, to build advanced data workflows.
	+ Expertise in using Looker Studio to create insightful and interactive data visualizations and dashboards.
	+ Optimized MongoDB queries for performance improvements.
	+ Extensive experience in developing Kafka producers and Kafka consumers for streaming millions of events per minute on streaming data using PySpark, Python & Spark Streaming.
	+ Integrated GCP Cloud Healthcare API for secure, HIPAA-compliant storage and management of healthcare data, enabling interoperability with other healthcare systems and services.
	+ Experienced in developing end to end automation using Selenium WebDriver/RC/IDE/Grid, Unittest/Pytest, Jenkins, GHERKIN/Cucumber, Robot, ALLURE reporting, RESTful API and PostMan.
	+ Conducted performance benchmarking and optimization of data structures in large-scale distributed systems.
	+ Developed ETL pipelines using Python to extract, transform, and load data from various sources.
	+ Developed complex SQL queries and stored procedures in Snowflake to support data analytics and reporting.
	+ Conducted thorough testing and debugging of DBT models to ensure high-quality data outputs and accurate analytics.
	+ Developed and implemented Historical and Incremental Loads using Databricks & Delta Lake run using ADF pipelines
	+ Hands - on experience in Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake .
	+ Working as data engineer and having strong background skills of big data technologies like Hive, Scala, Spark integrated with JAVA 8.
	+ Integrated Apache Iceberg with big data frameworks like Apache Spark, Flink, and Hive to enable seamless data processing and sharing.
	+ Experienced in deploying containerized data solutions on OpenShift, managing microservices and orchestrating data pipelines in a cloud-native environment.
	+ Monitored and tuned MongoDB database performance using tools like MMS and Ops Manager.
	+ Developed and optimized data processing jobs using Apache Spark, reducing ETL job runtimes significantly.
	+ Developed data pipelines using Talend Cloud ETL and AWS services like Lambda, S3.
	+ Used PySpark jobs to run on Kubernetes Cluster for faster data processing
	+ Created daily and weekly project status report as well as present power point slides to business and stakeholders regarding project progress.
	+ Worked with project manager for assigning task to different team members and track tasks as well as tracking deliverable.
	+ Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
	+ Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) BigQuery.
	+ Involved in Modeling and developing Business Requirements, Design process and providing Knowledge Transfer to the Support Teams.
	+ Partner with clients/technical groups to perform quality reviews on business requirements to determine if the requirements fulfill the demands of the end users.

**Client: DELOITTE, USA MAY 2021 – AUGUST 2022**

**Title: Sr. Data Engineer**

**Responsibilities:**

* + Worked closely with project stakeholders, business SME’s for gathering requirements and specifications for the applications and preparing business requirements document.
	+ Developed data pipelines using Talend Cloud ETL and AWS services like Lambda, S3.
	+ Experience in Kafka tool for Data testing.
	+ Implemented data sharing and collaboration across teams using Microsoft Fabric's OneLake shortlinks and unified datasets.
	+ Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters In Databricks, Managing the Machine Learning Lifecycle.
	+ Developed complete end to end Big-data processing in Hadoop eco system.
	+ Monitored and optimized data warehouse performance, resolving any issues related to data loading and processing.
	+ Experienced in developing and executing manual and automated tests in different platforms using Python, Pytest/Unittest/Robot and the Selenium library.
	+ Migrated on-premises environment in GCP (Google Cloud Platform)
	+ Automated failure handling and retry logic in Prefect to ensure high data pipeline uptime.
	+ Utilized Databricks' Delta Lake for managing and optimizing large-scale data lakes with ACID transactions and schema enforcement.
	+ Built versioned data pipelines using Apache Iceberg for time-travel queries and rollback capabilities, enhancing data reliability and debugging.
	+ Designed and implemented scalable ETL pipelines to ingest, transform, and share data across internal and external platforms.
	+ Used PySpark jobs to run on Kubernetes Cluster for faster data processing.
	+ Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery.
	+ Collaborated with DevOps teams to streamline CI/CD processes using OpenShift for automated deployment of ETL workflows, increasing system uptime and reliability.
	+ Enabled secure and governed data sharing by implementing fine-grained access controls and auditing on Iceberg tables.
	+ Developed and maintained shell scripts for system administration, data processing, and application deployment.
	+ Configured and maintained MongoDB replica sets and sharding for distributed databases.
	+ Optimized data storage and retrieval in Iceberg by leveraging its partitioning and indexing features.
	+ Experience with tree data structures (e.g., binary trees, AVL trees) to maintain sorted data and support quick search, insertion, and deletion operations.
	+ Designed and implemented enterprise data architecture strategies, ensuring alignment with business goals and IT standards.
	+ Conducted data profiling and data quality analysis using PowerCenter, ensuring data accuracy and consistency across multiple platforms.
	+ Experienced in developing and executing manual and automated tests in different platforms using Python, Pytest/Unittest/Robot and the Selenium library.
	+ Collaborated with cross-functional teams to define data-sharing requirements and build secure data exchange mechanisms.
	+ Experience in connecting Looker Studio to various data sources such as SQL databases, cloud storage, and APIs.
	+ Integrated Apache Flink with Apache Kafka for seamless real-time data streaming and processing.
	+ Skilled in configuring StreamSets pipelines to ingest, transform, and route data from diverse sources including databases, cloud storage, and real-time streams.
	+ Implemented complex Identity and Access Management (IAM) policies to segregate duties, enforce least privilege, and enhance security for GCP projects and resources.
	+ Utilized GCP Cloud Shell for remote command-line access to automate day-to-day operations, infrastructure provisioning, and debugging tasks.
	+ Secured sensitive data using GCP Secret Manager to centrally manage, access, and rotate secrets (API keys, passwords, etc.) across services and applications.
	+ Configured and managed Apache Kafka clusters for real-time data ingestion and streaming analytics.
	+ Implemented real-time monitoring and alerting for data workflows in Dagster, improving operational visibility.
	+ Developed and optimized complex ETL pipelines in Scala, leveraging its functional programming features for cleaner and more efficient code.
	+ Utilized MongoDB Aggregation Framework for complex data analysis.
	+ Developed and optimized complex SQL queries, stored procedures, and functions for Oracle databases.
	+ Developed multithreaded programs using Core Java to measure system performance.
	+ Collaborated with data analysts and business stakeholders to understand requirements and translate them into effective DBT models.
	+ Configured Iceberg catalogs with AWS Glue, Hive Metastore, and other backends to support dynamic schema evolution and multi-cloud environments.
	+ Developed and implemented coding standards and best practices to ensure high code quality and maintainability using Apache Spark with Scala and Pyspark, Hadoop.
	+ Experience with various shell scripting languages including Bash, KornShell, and C Shell.
	+ Familiar with deploying Python applications in cloud environments such as AWS, Azure, or GCP.
	+ Experienced in reverse engineering existing databases to create comprehensive data models in Erwin.
	+ Monitor and maintain CI/CD pipelines in Azure environment for data load from Lake to DBX and from DBX to SQL DW
	+ Working Knowledge of Delta Lake
	+ Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) BigQuery.
	+ Developed and maintained data pipelines leveraging efficient data structures to handle high-throughput data ingestion and processing.
	+ Built reports for monitoring data loads into GCP and drive reliability at the site level.
	+ Experienced in creating complex data pipelines with StreamSets, utilizing built-in processors, transformations, and connectors to handle various data formats and schemas.
	+ Integrated Prefect with cloud storage and database systems for seamless data management.
	+ Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
	+ Spearheaded the integration of new data sources, expanding the data warehouse to support additional business functions.
	+ Ensured compliance with data privacy regulations (e.g., GDPR, CCPA) by implementing secure and auditable data-sharing processes.
	+ Worked on NoSQL databases like MongoDB, Document DB and Graph Databases like neo4j .
	+ Hands - on experience in Azure Analytics Services - Azure Data Lake Store (ADLS), Azure Data Lake Analytics (ADLA), Azure SQL DW, Azure Data Factory (ADF), Azure Data Bricks (ADB) etc.
	+ Conducted stakeholder interviews and workshops to gather business requirements, analyzed business data to identify trends and insights, and developed reports and presentations to communicate findings and recommendations to senior management.
	+ Collaborated with cross-functional teams, including IT, operations, and finance, to ensure project success.
	+ Demonstrated strong problem-solving skills and a track record of delivering solutions that meet business requirements.
	+ Created PySpark scripts to encrypt raw data using hashing algorithms based on client-specified columns.
	+ Used Azure DeVops to document the requirements and followed Agile methodology.

**Client: Prokarma, Hyderabad, India NOVEMBER 2018 – MAY 2021**

**Title: Data Engineer**

* Collaborate on all stages of systems development lifecycle (SDLC), from requirements gathering to production releases.
* Developed data pipelines using Talend Cloud ETL and AWS services like Lambda, S3.
* Experience in Kafka tool for Data testing.
* Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters In Databricks, Managing the Machine Learning Lifecycle.
* Monitored and optimized data warehouse performance, resolving any issues related to data loading and processing.
* Experienced in developing and executing manual and automated tests in different platforms using Python, Pytest/Unittest/Robot and the Selenium library.
* Configured VPC Service Controls to define security perimeters around sensitive GCP resources, protecting data from unauthorized access and minimizing security risks.
* Automated failure handling and retry logic in Prefect to ensure high data pipeline uptime.
* Utilized Databricks' Delta Lake for managing and optimizing large-scale data lakes with ACID transactions and schema enforcement.
* Streamlined real-time data sharing by integrating Iceberg with streaming platforms like Apache Kafka or Flink.
* Developed and maintained shell scripts for system administration, data processing, and application deployment.
* Configured and maintained MongoDB replica sets and sharding for distributed databases.
* Optimized data storage and retrieval in Iceberg by leveraging its partitioning and indexing features.
* Experience with tree data structures (e.g., binary trees, AVL trees) to maintain sorted data and support quick search, insertion, and deletion operations.
* Designed and implemented enterprise data architecture strategies, ensuring alignment with business goals and IT standards.
* Designed and implemented role-based access controls to enable secure data sharing while maintaining governance standards.
* Conducted data profiling and data quality analysis using PowerCenter, ensuring data accuracy and consistency across multiple platforms.
* Review test cases and make sure they are in accordance with requirements and assist the overall testing process by activities such as Change and Defect Management.
* Managed testing cycles, reviewing test plan creation, development of scripts and co- ordination of user acceptance testing for software applications.

**Client: Microsoft, Hyderabad, INDIA (Data Engineering & Cloud Team) MARCH 2013 – MAY 2018**

**Title: Data Engineer/Software Engineer**

**Responsibilities:**

* SDLC collaboration: Involved in all stages from requirements gathering to production releases.
	+ Developed & Implemented a user-friendly UI based file format conversion; saved license cost
	+ Implemented NIFI architecture for data migration as a part of Hadoop & Cloud Team
	+ Worked on NoSQL databases like MongoDB, Document DB and Graph Databases like neo4j.
	+ Developed generic Talend Cloud ETL frameworks.
	+ Involved in project design and development using Java, Scala, Go, Hadoop, Spring, Apache Spark, NIFI, Airflow Technologies.
	+ Experience building distributed high-performance systems using Spark and Scala.
	+ Created AWS Glue crawlers for crawling the source data in S3 and RDS.
	+ Utilized Apache Iceberg to manage large-scale tabular data in a data lake, enabling efficient data querying and management.
	+ Optimized data-sharing performance by implementing partitioning, indexing, and caching strategies for distributed systems.
	+ Developed Tableau data visualization using Cross tabs, Heat maps, Bar Charts, Gantt charts, Waterfall charts, Scatter Plots, Geographic Maps, Pie Charts and Donut Charts.
	+ Documented DBT workflows and data models, providing clear and comprehensive guidelines for team members and stakeholders.
	+ Worked with Apache Airflow to schedule and monitor complex ETL workflows, ensuring data pipeline reliability..
	+ Implemented data security best practices in Snowflake, including role-based access control and data encryption.
	+ Experience with cloud-based data warehousing solutions (e.g., AWS Redshift, Google BigQuery, Snowflake).
	+ Develop overall Test Strategy, lead testing of all impacted applications/infrastructure and postproduction support activities.
	+ Implemented the Big Data solution using Hadoop, hive and Informatica to pull/load the data into the HDFS system.
	+ Experience building Data pipeline for Realtime streaming data and Data Analytics using Azure cloud components like Azure Data Factory, HDInsight (spark cluster), Azure ML Studio, Azure stream Analytics, Azure Blob Storage, Microsoft SQL DB, Neo4j (Graph DB).
	+ Wrote comprehensive documentation and provided training sessions for team members on best practices and advanced usage of Apache Flink.
	+ Created custom Python scripts for data migration, transformation, and loading tasks.
	+ Collaborating with cross-functional teams to gather requirements and translate them into detailed Erwin data models.
	+ Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark)
	+ Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop
	+ cluster on AWS EMR.
	+ Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
	+ Automated data workflows and scheduled jobs in Snowflake using Snowflake Tasks and Streams.
	+ Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
	+ Created multiple Apache NiFi custom templates with Spark to process data from different data sources and stored the Spark’s output into S3 buckets in parquet formats while implementing Amazon Lamda and Step functions to schedule the data flow
	+ Implemented Shell script to analyze the audit logs of hive database in real time while handling a team of 4 members
	+ Expertise in using JAVA, BIG DATA, HTML, DHTML, CSS, D3JS, JSON, Font awesome, JavaScript and Bootstrap, Neo4j and Keylines.
	+ Used PySpark jobs to run on Kubernetes Cluster for faster data processing.

# Academic projects (Business & CLOUD Services & CRM):

1. **Chicago Food inspection Analysis (Fall 2022)**

I acquired and organized information about food inspections in the area as part of the Chicago Food Inspections Analysis project. I developed an interactive dashboard using Tableau that allowed users to quickly grasp the data and spot areas that required more attention.

The dashboard provided a thorough overview of the regional food inspection process by allowing users to filter the data by risk category, inspection type, and location. The project gave me the opportunity to practice data collecting, analysis, and visualization while also using statistical methods to solve actual issues.

1. **Analysis of District Level Standardized Test Performance in Pennsylvania Public Schools (Spring 2023)**

In this project, I used JMP tool for data mining and Analysis. Performed many Machine learning **techniques to come up with a project report suggesting the school authorities to make minor changes in their curriculum to increase the level of standardized test performance of the students.** Silicon Valley Education Foundation (SVEF)

* Involved in upgrading to Salesforce Nonprofit Starter Pack (NPSP) 3.x Account Model conversion
* Worked with Visualforce Pages and Salesforce Sites.
* Regularly performed database deduping and cleanup procedures.
* Worked on Standard, Custom Objects and user Setup.
* Sales and Service Application
* Data Management
* App Builder Configuration.

Organization Set-Up and Workflow/Process Automation. (Both flows and Process builders)