Nilay Vyas

Data Engineer

Email: nvdata03@gmail.com

Add: Frisco, United States **Call: +1 (214)-764-9442**

LINKS

LinkedIn: https://www.linkedin.com/in/nilay-v/

PROFILE

Experienced Data Engineer with a proven track record of designing and implementing cutting-edge data pipelines that drive actionable insights. Over 8 years in the field, I've spearheaded projects that resulted in an increase in data processing efficiency, while consistently ensuring data accuracy and security. My strong analytical mind-set, expertise in ETL processes, and proficiency in Python, SQL, and Big Data technologies make me the ideal candidate to transform data into a strategic asset for your organization. Let's turn your data challenges into opportunities together.

EMPLOYMENT HISTORY

★ Senior Data Engineer at Blue think Inc. …… Oct 2021 - Present

- Over 8 years of experience in Data Engineering, Data Pipeline Design, Development, and Implementation.
- Expertise in designing star schema, snowflake schema for Data Warehouses, and ODS architecture and Familiarity with Erwin Data Modeller and ER Studio for data modelling.
- Strong proficiency in various Big Data tools and technologies, including Hadoop Ecosystem (MapReduce, HBase, Hive, Pig, Sqoop, Kafka, Oozie), Spark (PySpark, Spark SQL), and Airflow for orchestrating data pipelines. Hands-on experience with Google Cloud components and client libraries.
- Extensive use of Python for data analysis, data manipulation, and scripting. Familiarity with Python libraries such as PySpark, Pytest, Pymongo, cxOracle, PyExcel, Boto3, Psycopg, embedPy, NumPy, and Beautiful Soup.
- Proficiency in working with Amazon Web Services (AWS) Cloud Platform, including services like EC2, S3, RDS, VPC, Lambda, Athena, EMR, Redshift, DynamoDB,

CloudFront, CloudWatch, Route 53, etc. Also experienced in Google Cloud Platform (GCP) services.

- Experience in developing ETL workflows and data pipelines using tools like SSIS, Informatica, and Talend. Knowledge of data extraction, transformation, and loading from heterogeneous sources.
- Data Analysis and Data Profiling: Skilled in data analysis, data profiling, data integration, data migration, and metadata management. Proficient in SQL across multiple dialects, including MySQL, PostgreSQL, Redshift, SQL Server, and Oracle.

★ Data Engineer & Analyst at BT IT Consulting……June 2015 – Sep 2021

- Developed an enterprise data model that integrated data from multiple sources and enabled consistent data access across the organization.
- Implemented solutions for ingesting data from various sources and processing the Data- Confidential -Rest utilizing Big Data technologies such as Hadoop, Map Reduce Frameworks, HBase, and Hive with Cloud Architecture.
- Worked on AWS, implementing solutions using services like (EC2, S3, RDS, VPC, and Lambda).
- Experience in Data Analysis, Data Profiling, Data Integration, Migration, Data governance and Metadata Management, Master Data Management and Configuration Management.
- Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using Python.
- Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)
- Extensively used Python Libraries PySpark, Pytest, Pymongo, cxOracle, PyExcel, Boto3, Psycopg, embedPy, NumPy and Beautiful Soup.

EDUCATION

★ Florida Institute of Technology..... July 2013 – June 2015

Masters in Engineering

★ Gujarat University..... August 2007 – June 2011

Bachelors in engineering

SKILLS

Data Warehousing:

- Design and implementation of data warehouse solutions using platforms such as Amazon Redshift, Google BigQuery, and Snowflake.
- Expertise in creating star and snowflake schemas for efficient data storage and retrieval.

ETL (Extract, Transform, Load):

- Proficient in ETL processes using tools like Apache NiFi, Apache Airflow, and Talend.
- Data extraction from various sources, transformation, and loading into data warehouses.

Programming Languages:

- Strong programming skills in Python for building data pipelines and automating data processes.
- SQL for data querying, optimization, and data modeling.

Big Data Technologies:

- Experience with Apache Hadoop ecosystem components like HDFS, MapReduce, and Hive.
- Proficient in Apache Spark for large-scale data processing and analysis.

Database Management:

- Database administration and optimization of relational databases such as PostgreSQL, MySQL, and Microsoft SQL Server.
- NoSQL database expertise, including MongoDB, Cassandra, and DynamoDB.

Data Modeling:

- Designing and maintaining data models using tools like Erwin, Lucidchart, or custom SQL scripts.
- Knowledge of dimensional modeling and data modeling best practices.

Data Integration:

- Integration of data from diverse sources, including REST APIs, JSON, XML, and flat files.
- Data synchronization and replication techniques.

Cloud Services:

- Hands-on experience with cloud platforms like AWS, Azure, and Google Cloud.
- Setting up and managing data storage, compute resources, and serverless architectures in the cloud.

Data Quality and Governance:

- Implementing data quality checks and data governance policies to ensure data accuracy and compliance.
- Proficient in data profiling and data cleansing techniques.

Version Control:

- Utilizing Git for version control and collaborative development of data engineering code.

DevOps and Automation:

- Experience with containerization tools like Docker and container orchestration using Kubernetes.
- Continuous integration and continuous deployment (CI/CD) pipeline setup for data workflows.

Monitoring and Logging:

- Implementing monitoring and logging solutions using tools like Prometheus, Grafana, ELK Stack (Elasticsearch, Logstash, Kibana).

Collaboration and Documentation:

- Effective communication skills and the ability to work in cross-functional teams.
- Documentation of data pipelines, workflows, and best practices.

PROJECTS

★ Project - **Project-AI (development)**

Description- Created pipeline for client using AWS Glue services, in this project created the Crawler for creating the schema from database in Datacatlog and created the job and trigger at the time.

Responsibilities:

- Designing the Pipeline: Determine the data sources, transformations, and destination for your pipeline. This includes identifying the required AWS Glue components such as crawlers, jobs, and triggers.
- Setting up Data Sources: Configure AWS Glue to connect to your data sources, which can include various databases, data lakes, or streaming services. Configure the necessary access permissions and connectivity options, Worked on the Pandas and Lambda.
- Defining Data Transformations: Create and configure AWS Glue jobs to perform the required transformations on your data. This may involve data cleaning, data enrichment, or data aggregation tasks, depending on your specific pipeline requirements.
- Used Python for XML, JSON processing, data exchange and business logic implementation.
- Building ETL Workflows: Use AWS Glue to define the order and dependencies of your data transformations. This includes creating ETL workflows or DAGs (Directed Acyclic Graphs) to ensure data is processed in the correct sequence.
- Implemented Big Data tools like Spark using Python and utilizing Data frames and Spark SQL API for faster processing of data and worked on extensible framework for building high performance batch and interactive data processing application on hive.

Environment: Python, Django, AGILE/SCRUM methodologies, MySQL, DataStage, Netezza, E3 Framework, Unix scripting, Hadoop 3.0, HBase 1.2, Hive 2.3, AWS, EC2, S3, RDS, VPC, MySQL, Redshift, Sqoop, HDFS, Spark, ETL.

★ Project - Health channel

Description- In this project, we have implemented a robust data management and automation solution leveraging Google Cloud Platform (GCP) services. Our goal was to streamline data workflows, improve data accessibility, and enhance data-driven decision-making.

Responsibilities:

- Develop, deploy, and maintain data pipelines using Google Dataflow to extract, transform, and load (ETL) data.
- Implement efficient data ingestion and integration processes from various sources into BigQuery.
- Design and optimize BigQuery datasets, tables, and schemas for efficient querying and storage.
- Create and manage BigQuery views and stored procedures to facilitate data access and transformation.
- Develop serverless functions using Google Cloud Functions for automation and integration purposes
- Implement event-driven processes and data triggers using cloud functions.

Environment-, Big query,store proc,view,data prep, dataflow, GCP function.

★ Project- MCE

Description: our primary objective is to streamline data operations, enhance data reliability, and empower data-driven decision-making by integrating multiple data sources and technologies. We are harnessing the power of Apache Airflow, Snowflake database, PySpark, SFTP server connections, and AWS S3 to achieve these goals.

Responsibilities:

- Design, develop, and maintain data pipelines using Apache Airflow to automate and schedule data workflows.
- Extract, transform, and load (ETL) data between different systems, including Snowflake, S3, and SFTP servers.
- Connect to Snowflake databases to extract and load data, ensuring data accuracy, reliability, and performance.
- Optimize SQL queries and data transfer processes within Snowflake.
- Utilize PySpark for data processing, including data cleansing, transformation, and aggregation.
- Establish secure connections to SFTP servers to exchange data with external partners or systems.
- Integrate with Amazon S3 for storing, retrieving, and managing data objects.

Environment: Hadoop/Big Data Technologies: Spark-Python, Kafka, Spark Streaming, Mlib, Sqoop, Hbase, HDFS, Map Reduce, Pig, Hive, AWS Glue ,Zeppelin(Distributions Data Bricks, Horton works and Cloudera), Cassandra, HBase, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, JDBC, Apache, Shell Scripting, Pandas, Lambda.

★ Project -Merchant (Leanscale)

Description: Data scraping of web site and order automation.

Responsibilities:

- Created Script for Amazon, Ali expresses, Walmart, Asos.com to Scrap data and store the data inside MongoDB database.
- Used Python for XML, JSON processing, data exchange and business logic implementation.
- Expertise in developing data driven applications using Python 2.7, Python 3.0 on PyCharm and Anaconda Spyder IDE's.
- Created Script for Order automation for these websites.
- Developed Script on AWS server with MongoDB database.

Environment- Python, Django, Bootstrap, Html, MySQL, JavaScript, CSS, Integromate, Nexmo ,Mtalkz, Postmark, GCP function, Nginx ,Gunicorn , AWS.