**Vinay B.**

Cloud Data Engineer

Detroit, Michigan **•** 4244284629 **•** [vinaydataengineer23@gmail.com](mailto:vinaydataengineer23@gmail.com) [**•**](https://www.linkedin.com/in/vinaydatasaur)[linkedin.com/in/vinaydatasaur](https://www.linkedin.com/in/vinaydatasaur)

**SUMMARY**

* **Data Engineer** with **10 years of** experience, leveraged **Big** **Data** **technologies** and Native Cloud Services, **AWS, Azure, GCP** to deliver data-centered analytics solutions to sectors such as finance, healthcare, and Pharma.
* Utilized **ETL optimization, data pipeline creation using Alteryx, SQL, Python, and Data Warehousing**, leading to a 40% acceleration in processing, a 90% decrease in latency, and a 60% enhancement in fraud detection within a year.
* Spearheaded the **establishment of data governance**, fostering cross-team collaboration and alignment of analytics with objectives, resulting in a 50% improvement in marketing analytics and a 90% reduction in data breaches.
* Extensive experience in end-to-end **Big Data Application Design**, with expertise in Hadoop Ecosystem components like Apache Hadoop Map Reduce, Spark, HDFS, HIVE, PIG, HBase, Zookeeper, Sqoop, Oozie, Kafka, Python, and Scala.
* **Strong expertise in building scalable applications** using various programming languages (Java, Scala, and Python).
* **Expertise in formulating Apache Spark programs using Python (PySpark)** for topology construction and data manipulation using Datasets and Dataframe.
* 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-Premises databases to Azure Data Lake store using Azure Data factory.
* Used AWS Services like **S3, Athena, RDS, DynamoDB, Redshift, Lambda, Glue, Ec2, EMR, Sage Maker, Step function, CloudWatch, SNS, IAM Role, VPC, Kinesis, QuickSight**, etc.,
* Used GCP Services like **Cloud Storage, BigQuery, Cloud Function, Compute Engine, Data Flow, Data Proc, Cloud Composer, Cloud Run, Cloud Monitoring,** etc.,
* Coordinated with **data architects for logical models’ translation** into an optimal physical model that performs efficiently under various load profiles, aiding in Market business initiatives.
* **Analyzed and visualized data with R, Python and created dashboards** using tools like Tableau, Power BI, revealing trends and relationships between different pieces of data.
* **Streamlined data pipelines using AWS services** including EC2, S3, Redshift, Glue, Lambda functions, Step functions, Cloud Watch, SNS, DynamoDB, SQS.
* Optimized **Snowflake Schemas by normalizing the dimension tables** and creating a Sub Dimension named Demographic as a subset to the Customer Dimension.
* Holds good knowledge of **Data Marts, OLAP, Dimensional Data Modeling with Ralph Kimball Methodology** (Star Schema Modeling, Snow-Flake Modeling for FACT and Dimensions Tables) using Analysis Services.
* Experienced working with various **Hadoop Distributions (Cloudera, Hortonworks, Amazon EMR)** to fully implement and leverage various Hadoop services.
* Streamlined Mappings design and **enhanced performance tuning for slowly changing Dimension Tables and Fact tables**.
* Orchestrated the **validation of ETL processes between multiple databases** like Oracle, SQL Server, Hive, and MongoDB using Python, leading to robust regression scripts.
* Expertise in using **Docker and setting up ELK with Docker and Docker-Compose**. Actively involved in deployments on Docker using Kubernetes.
* Delivered full life cycle support to **logical/physical database design, schema management, and deployment**.
* Established **Docker and Kubernetes for the runtime environment of the CI/CD system**, aiding in build, test, and deploy.
* Adopted **analytical applications like R, SPSS, Rattle, Python** to identify trends and data relationships, translating analytical findings into risk management and marketing strategies.
* Took charge of the **development and deployment of enterprise-based applications using major Hadoop ecosystem components**.
* Fostered efficiency of **existing algorithms using SparkCore, Spark SQL, Spark MLlib, DataFrame, Pair RDD, and Spark YARN**.
* **Skilled in SQL across several dialects**, including MySQL, PostgreSQL, Redshift, SQL Server, and Oracle.
* Showcased **excellent communication skills**, successfully working in a fast-paced multitasking environment, both independently and in a collaborative team.
* Proven ability to work effectively in **cross-functional team environments**, demonstrating excellent communication, and interpersonal skills.

# CERTIFICATIONS

Alteryx Designer Core/ Designer Cloud Core

Amazon Web Services Cloud Practitioner

Microsoft Azure Fundamentals

Alteryx Machine Learning

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Programming Languages | Python, Shell Scripting, Scala, PL-SQL, TSQL, DAX |
| Big Data Technologies | Hadoop, Spark, MapReduce, Hive, NiFi, Cassandra, Flink, Kafka |
| Cloud Platforms | AWS (S3, Kinesis, EMR, RDS, Lambda, QuickSight, Fargate, Glue,  EC2, ECS), Azure (HDInsight, SQL Database, Blob Storage, Synapse Analytics, Data Lake Storage Gen2, Cosmos DB, Event Hubs, Monitor) |
| Databases | MariaDB, MySQL, Teradata, MongoDB, PostgreSQL, SAP HANA |
| Data Warehousing & ETL Tools | Alteryx, Informatica, SSIS, Azure Data Factory, Snowflake, Google BigQuery, AWS Kinesis, AWS Glue |
| Data Visualization | AWS QuickSight, Power BI, Tableau, QlikView, Seaborn, Matplotlib, Plotly |
| Machine Learning Libraries | TensorFlow, PyTorch, XGBoost, scikit-learn, Pandas, NumPy, SciPy, Statsmodels, NLTK, AWS SageMaker |
| DevOps and Version Control | Docker, Kubernetes, AWS CodePipeline, Jenkins, Terraform, GitHub, Azure DevOps |
| Workflow Management | Apache Airflow, AWS Step Functions |
| Project Management Tools | Jira, Agile/Scrum Methodology, Visio |
| Other Tools | GraphQL, REST API, Prometheus, Kibana, Neptune, Kubeflow, EPIC |

**WORK EXPERIENCE**

**EPAM Systems •** Sr Data Engineer **•** RemoteSep 2021 – Present

* Optimized data availability and reduced processing time by **20%** through the construction of an ETL process using **AWS Glue**.
* Achieved a **25%** decrease in storage costs and enhanced data scalability by designing and implementing a data lake solution with **Amazon S3**.
* Collaborated on data encryption policies, resulting in a **99.99%** reduction in data breaches and ensuring regulatory compliance.
* Improved query execution speed by **30%** and expedited report generation by developing optimized SQL queries.
* Orchestrated efficient data ingestion into the Redshift data warehouse by leveraging **Spark** scripts for data cleansing and analysis.
* Facilitated seamless data management and transfer between local and Amazon S3 through the creation of **AWS** **Lambda** functions for file operations and data upload.
* Enabled real-time data streaming, allowing prompt adaptation to market changes, using **Amazon Kinesis**.
* Ensured uninterrupted business operations and **99.9%** data availability by implementing disaster recovery solutions with **AWS CloudEndure**.
* Improved system response times by **30%** during high data loads through the utilization of **AWS Elastic Beanstalk**.
* Streamlined pipelines and reduced deployment time by **40%** through the application of CI/CD techniques, **AWS CloudFormation**, and **Terraform**.
* Translated client requirements into functional specifications and facilitated onboarding onto the Hadoop ecosystem.
* Ensured seamless data integration and management by extracting and updating data into HDFS using **Sqoop** import and export.
* Designed and developed **Flink** pipelines for real-time data consumption and transformation from **Kafka**, facilitating efficient data processing and analysis.
* Streamlined data processing and querying through the development of AWS Lambda functions for generating a serverless data pipeline, utilizing **Glue Catalog** and **Athena**.
* Enhanced data processing capabilities by constructing an on-demand, secure EMR launcher with custom Spark submit steps, integrating **S3 Event**, **SNS**, **KMS**, and **Lambda function**.
* Implemented AWS data pipelines incorporating various services such as **VPC**, **EC2**, **S3**, **Auto Scaling Groups (ASG)**, **EBS**, **Snowflake**, **IAM**, **CloudFormation**, **Route 53**, **CloudWatch**, and **CloudFront**.
* Enabled advanced data analysis and modeling through the development of MapReduce/Spark Python modules for machine learning and predictive analytics.
* Managed Hive tables implemented partitioning strategies, and integrated **Spark** for interactive queries and integration with popular **NoSQL** databases.
* Implemented end-to-end CI/CD pipelines, integrating SCM (Git) with Gradle for automated testing, and deploying applications using Jenkins, Docker, and Kubernetes for efficient software delivery.
* Ensured reliable data streaming and processing by managing data ingestion from various sources through Kafka, implementing custom producers and consumers.
* Enabled streamlined data analysis and migration processes through the development and execution of **Databricks** notebooks using SQL, Python, and automated notebooks.
* Leveraged the power of Spark to improve performance and scalability by migrating MapReduce jobs to Spark jobs.

#### **JPMorgan Chase & Co •** Data Engineer **•** Wilmington, DE Jul 2019 -Aug 2021

* Collaborated with various teams to create conceptual, logical, and physical data models for **Business Intelligence reporting databases**, resulting in a **30% improvement** in reporting accuracy.
* Developed advanced research techniques to investigate and resolve complex compatibility and interface design considerations, reducing system failures by **25%**.
* Built, edited, and reviewed complex SQL procedures using **DB2** and **SQL Server**, resulting in a **20% increase** in database query efficiency.
* Performed **Root Cause Analysis** on existing ETL processes, resulting in a **30% reduction** in issues with supplied data sources.
* Developed complex mappings using **Informatica cloud integration services**, leading to an enhanced data integration efficiency by **30%**.
* Developed dashboards in **Microsoft Power BI**, aiding in efficient data interpretation and enhancing business decision making.
* Hands-on experience with building data pipelines in **Python**, **Pyspark**, **Hadoop**.
* Analyze data quality issues through exploratory data analysis using **SQL** and **Pandas**.
* Managing, Configuring, and creating **AWS services** using **Boto3**.
* Transforming, Ingesting, and automating terabytes of real-time into **parquet format** utilizing **Kinesis**, **Lambda**, and **Apache Airflow**.
* Working on real-time data streaming using **Amazon Kinesis**.
* Working on Ingesting, Transforming, and loading the data to **S3**, **Amazon Elastic Search**, and **Amazon Redshift** using **Kinesis Data Firehose**.
* Performed real-time aggregations and filtering process of streaming data using **Kinesis Data Analysis**.
* Working on creating automation scripts for leveraging various python libraries to perform accuracy checks from various sources to target databases.
* Working on **Authentication** and semantics of invoking **AWS Service APIs**.
* Created and Ingested data into **S3 Bucket**. Created modular components for AWS S3 connection, data reads.
* Executing interactive analytics like cleansing, and validating data on S3 Bucket using **AWS Athena**.
* Extracting the data from **PostgreSQL** and **AWS Redshift** into **Hadoop Distributed File System** utilizing **Kinesis**.
* Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as **Amazon S3** and **Amazon DynamoDB**.
* Using **Glue Crawler** to crawl the data, working on developing **Glue ETL jobs** in the notebook, Deploy and Running the Glue ETL jobs.
* Develop **AWS Glue Data Pipelines** for ingesting, manipulating, and storing the data while interacting with different services in AWS.
* Orchestrating the jobs using **AWS Step Function**, and **AWS Lambda**. Configured various performance metrics using **AWS CloudWatch**.
* Developing custom python scripts for automating the **ETL Data pipeline** and **DAG workflows** using **Airflow**.
* Hands-on experience in **CI/CD**, **Cloud Formation**, **EC2**.
* Implemented **AWS Step Function** to automate and orchestrate the **Amazon SageMaker** for predicting the user.

#### **Schneider Electric •** Data Engineer • Franklin, TN Sep 2018 - Jun 2019

* Constructed Alteryx custom-built input connectors to ingest and analyze large datasets from RDBMS to **Azure SQL Data Warehouse** for customer segmentation.
* Designed **ETL workflows** with Alteryx to transfer data from TSV files/Oracle database into **Azure Data Lake Storage**, decreasing manual effort by 40%.
* Managed over 80 schedules and 120 workflows on **Alteryx Server**, ensuring smooth operation of different business units.
* Built a **real-time data processing pipeline** using Alteryx with clustering techniques that improved customer engagement by 25%.
* Extracted data from ISU and CRM and designed a dashboard for service orders using Alteryx Workflows with the help of **DAX coding**.
* Utilized **real-time streaming data** in the logistic and device tracking team, leveraging **Azure Event Hub** (AMQP and Kafka) for data extraction and processing.
* Used **Azure IoT Hub** and **Kafka** to capture bi-directional communication between devices and scalable message queuing for data ingestion.
* Worked on **Kafka** for message queuing and streaming process application for handling and routing messages.
* Developed **Spark programs** using Scala to compare the performance of Spark with Hive and SparkSQL.
* Devised **PL/SQL Stored Procedures, Views, and Triggers**, enhancing **data quality and integrity** by 20%.
* Excelled in T-SQL query development and analysis, creating accurate data extracts that improved business decision-making.
* Created **Azure Data Pipeline** using python packages like Azure Management Resources, Data Factory, Azure Identity.
* Utilized various libraries from Data Factory like **Resource Management client**, time delta.
* Created a data pipeline using **Copy activity** by passing the parameters to the pipeline and the data flow.
* Worked on parsing CSV, JSON, and Avro files for **Azure Stream Analytics**.
* Performed **Data Cleansing**, **Data Reduction**, and **Data storage** on SQL Data warehouse using Azure Stream Analytics running on IoT Edge.
* Created custom Python scripts for **training and testing data** from Azure Storage.
* Developed advanced SQL queries for **transforming and managing the data**.
* Managed to store the data collected on Azure Storage Blob Containers or Azure Data Lake store and **Cosmos DB** for faster response time.
* Implemented **Orchestrating** the data using the **Azure Data Factory**.
* Created **Airflow DAGs** for the orchestration and triggering of the jobs.
* Utilized **Azure Monitor logs** and PowerShell using the Azure portal for monitoring activities and pipelines.

**IQVIA**• Data Engineer • Durham, NC May 2017 - Aug 2018

* Configured **Spark Streaming** in Python to receive real-time data from **Kafka** and store it into **HDFS**, ensuring timely availability of data for analysis.
* Collaborated with a cross-functional team of scientists and engineers to design and deliver **data infrastructure** solutions, significantly improving the informatics organization's efficiency and data-driven decision-making capabilities.
* Successfully troubleshooted and debugged **Hadoop ecosystem** run-time issues, ensuring smooth data processing operations.
* Developed **PySpark scripts** in Python for automated extraction of required data from data sets and storing it on **HDFS**, streamlining data accessibility and availability.
* Leveraged exceptional **communication skills** to influence stakeholders and collaborate effectively across teams, fostering a collaborative and informed work environment.
* Orchestrated the automation of scripts and workflows using **Apache Airflow** and shell scripting, resulting in daily execution in a production environment and improving operational efficiency by 40%.
* Demonstrated strong **analytical and problem-solving skills** in developing and implementing data mapping procedures, ETL processes, and data analysis using **R programming**, ensuring accurate and insightful analysis of data.
* Played a key role in the migration and integration of data from various sources, including files and RDBMS, ensuring data integrity and availability for analysis.
* Led the development and maintenance of **data pipelines** using **NiFi**, ensuring seamless data ingestion into the Data Lake and efficient data flow.
* Developed and implemented **Spark streaming applications** to consume real-time data from **Kafka**, perform transformations, and generate insights, enabling real-time data analysis.
* Collaborated with the ETL/Informatica team to define relevant attributes, dimensions, hierarchies, and metrics, ensuring accurate and comprehensive reporting in **Cognos**.
* Developed and published interactive reports and dashboards using **Tableau**, providing stakeholders with intuitive visualizations, and enabling data-driven decision making.
* Demonstrated **leadership skills** in managing projects, coordinating cross-functional teams, and taking ownership of deliverables, ensuring successful project completion and client satisfaction.
* Actively participated in team collaboration and achieved common goals by effectively working with team members, contributing to a positive and productive work environment.

# Eli Lilly & Company • ETL Developer • Indianapolis, Indiana Jan 2015 - Apr 2017

* **Defined** end-to-end clinical data strategy, leading to a 30% improvement in data flow and normalization for **Data & Analytics standards**.
* Transformed business requirements into **data warehousing solutions** via **Informatica Designer**, enhancing data accessibility and understanding by 25%.
* **Engineered ETL processes** utilizing Informatica 9.1.1, ensuring a 20% increase in the efficiency of **data integration** from various sources.
* Utilized an array of transformations for enhanced data management, including Lookup, Update Strategy, Expression, Aggregator, Filter, Stored Procedures, and Joiner.
* **Constructed UNIX shell scripts** for vendor data extraction and integration into the Informatica environment, achieving consistent, consumable data.
* Generated **Reports using Charts, Gauges, Tables, matrix**, enabling effective data representation and analytics.
* Orchestrated the processing of web services data through Informatica, transforming and loading XML responses into Oracle staging tables.
* Developed **ETL pipelines using Python and Snowflake SnowSQL**, fostering data interchange and consistency.
* **Facilitated data analysis** and resolved ad-hoc requests in liaison with business analysts and clients, boosting portfolio delivery by 15%.
* **Optimized Informatica ETL mappings** using caches, SQL query overrides, and parameters, leading to a 25% increase in data processing speed.
* Developed Teradata processes using Shell Scripting and RDBMS utilities, demonstrating strong initiative and problem-solving skills.
* Automated workflows and populated parameter files using pmcmd commands on command prompt and UNIX Shell scripts.
* **Spearheaded** transformations development with varied file formats, promoting **diverse data modality** and seamless **data flow**.
* Ensured code migration from Dev to Test to Prod environment, while crafting comprehensive techno-functional documentations and test cases for SDLC compliance.

# Airports Authority of India • Data Reporting Analyst • Coimbatore, India Mar 2011 - Dec 2012

# Streamlined data warehousing efforts by defining and analyzing requirements, facilitating 10% improvement in data storage and analysis capabilities.

* Revolutionized data processing capabilities by implementing complex business logic in **Informatica** mappings, sessions, and workflows, **increasing** data analysis accuracy by 35%.
* Boostedunderstanding **of data quality and structure** by conducting **data profiling and data analysis** on Oracle and flat files.
* Established a **data archival strategy for historical flight data** using **Oracle 11g**, resulting in **improved database performance** and enabling **efficient analysis of long-term trends** in flight patterns and delays.
* Delivered technical specifications guiding the **ETL** development and testing, improving efficiency by 30%.
* Developed proficiency in **Data Modeling**, including Dimensional (Star Schema) and Relational Normalized modeling including 3NF.
* **Improved database performance by 20%** by establishing a data archival strategy for historical flight data using **Oracle 11g**, enabling efficient trend analysis.
* Ensured timely and accurate data processing by scheduling and monitoring workflows using **Informatica Workflow Manager and Workflow Monitor**.
* Automated **data processing operations** by collaborating with **Autosys Scheduler**, resulting in improved efficiency.
* Acquired strong motivation, aptitude, high attention to detail, excellent communication skills, creative problem-solving, flexibility, and proactive attitude to learn newer tools/technologies.

# EDUCATION

**Bachelor of Science in Electronics Engineering Technology** Jun 2008 - Jun 2012

Sreenidhi-Vaughn **•** Hyderabad, Telangana

**Master of Science in Information Technology Management** Jan 2013 - Dec 2014

Lamar University **•** Beaumont, Texas