**Sr Data Engineer**

**Name: Viritha D**

**Ph No:** **8152937333**

**Email: viritha226@gmail.com**

# Professional Summary:

* Having around 10+ years of professional experience in IT, working with various Legacy Database systems, which include work experience in Big Data technologies as well.
* Good experience in understanding the architecting, designing, and operation of large-scale data and analytics solutions on Snowﬂake Cloud Data Warehouse.
* Experience in 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.
* Experience in Requirement gathering, System analysis, handling business and technical issues & communicating with both business and technical users.
* Hands-on experience in complete Software Development Life Cycle SDLC for the projects using methodologies like
* Agile and hybrid methods.
* Stay up-to-date on the latest advancements in data engineering tools and technologies (e.g., Apache Spark, Airflow, Snowflake, Data Bricks)
* Experience in analyzing data using Big Data Ecosystems including HDFS, Hive, HBase, Zookeeper, PIG, Sqoop, and Flume.
* Knowledge and working experience with big data tools like Hadoop, Azure Data Lake, and AWS Redshift.
* Good understanding of Apache Airﬂow.
* Designed, implemented, and maintained end-to-end data pipelines on Databricks to ingest, transform, and load large volumes of structured and unstructured data.
* Leveraged Databricks Delta Lake for managing and optimizing data storage, ensuring reliability, consistency, and ACID compliance.
* Experience in workﬂow scheduling with Airﬂow, AWS Data Pipelines, Azure, SSIS, etc.
* Experience in Migrating databases 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.
* Good working knowledge of Snowflake and Teradata databases.
* Good understanding of Big Data Hadoop and Yarn architecture along with various Hadoop Demons such as Job Tracker, Task Tracker, Name Node, Data Node, Resource/Cluster Manager, and Ka a (distributed stream-processing).
* Experience in Text Analytics, Data Mining solutions to various business problems and generating data visualizations
* using SAS and **Python**.
* Experience in Developing Spark applications using Spark - **SQL** in **Databricks** for data extraction, transformation, and aggregation from multiple ﬁle formats for analyzing & transforming the data to uncover insights into customer usage patterns.
* Hands-on experience building **PySpark**, Spark Java, and Scala applications for batch and stream processing involving Transformations, Actions, Spark SQL queries on RDDs, and Data frames.
* Strong experience writing, troubleshooting, and optimizing Spark scripts using Python, and Scala.
* Experience with PostgreSQL features such as ACID transactions, row-level locking, and replication.
* Good understanding of Spark Architecture including Spark Core, Spark **SQL**, Data Frames, Spark Streaming, Driver Node, Worker Node, Stages, Executors, and Tasks.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), Data Factory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes.
* Implemented security best practices in Azure Databricks, including authentication, authorization, and data encryption, ensuring data security and compliance with Azure data governance standards.
* Involved in building Data Models and Dimensional Modeling with 3NF, Star, and Snowflake schemas for OLAP and Operational data store (ODS) applications.
* Collaborated with cross-functional teams to design and implement data solutions using Azure Databricks, showcasing teamwork and communication skills in a collaborative Azure environment.
* Stayed up-to-date with the latest **Azure** **Databricks** features, updates, and best practices, continuously improving skills and knowledge in **Azure** **Databricks** and data processing technologies in the **Azure** ecosystem.
* Strong experience and knowledge of No**SQL** databases such as MongoDB and Cassandra.
* Experience in development and support knowledge of Oracle, **SQL**, PL/**SQL**, and T-**SQL** queries.
* Experienced in conﬁguring and administering the Hadoop Cluster using major Hadoop Distributions like Apache Hadoop and Cloudera.
* Experience integrating metadata using AbInitio's Metadata Hub and Express
* Experienced in building highly scalable Big-data solutions using NoSQL column-oriented databases like Cassandra, MongoDB, and HBase by integrating them with Hadoop Cluster
* Implemented advanced Spark optimizations, such as partitioning, caching, and parallelism, to improve query performance and resource utilization.
* Experience with using PostgreSQL in a cloud environment, such as AWS, Azure, or GCP.
* Experience with Jira, Confluence, and Rally for project management and Oozie, Airflow scheduling tools
* Solid experience creating cloud-based solutions and architecture using Amazon Web services (Amazon EC2, Amazon S3, Amazon RDS, EMR, Glue) and Microsoft **Azure**.
* Experienced in Technical consulting and end-to-end delivery with architecture, data modeling, data governance and
* design - development - implementation of solutions.
* Experience with designing, developing, and maintaining data pipelines using PostgreSQL.
* Experience in Big Data Hadoop Ecosystem in ingestion, storage, querying, processing, and analysis of big data.
* Extensive working experience in an agile environment using a CI/CD model.
* Experienced in Strong scripting skills in Python, Scala, and UNIX shell.
* Extensive experience working with structured data using Spark **SQL**, Data frames, Hive QL, optimizing queries, and incorporating complex UDFs in business logic.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data & Hadoop Ecosystem** | Hadoop 3.3/3.0, Hive 2.3, Solr 7.2, Apache Flume 1.8, Sqoop 1.4, Ka a 1.0.1, Oozie 4.3, Hue, Cloudera Manager, Stream sets |
| **Cloud Technologies:** | AWS, Glue, EC2, EC3, EMR, Redshift & MS Azure, Snowﬂake |
| **Programming Languages:** | Python, Scala, SQL, Java, C/C++, Shell Scripting |
| **Data Modeling Tools:** | Erwin R9.7, ER Studio v16 |
| **Packages**: | Microsoft Oﬃce 2019, Microsoft Project, SAP and Microsoft Visio 2019, Share point Portal Server |
| **RDBMS / NoSQL Databases:** | Oracle 19c, Teradata R15, MS **SQL** Server 2019, Cosmos DB, Cassandra 3.11, HBase 1.2, |
| **Testing and defect tracking Tools:** | HP/Mercury, Quality Center, Win Runner, MS Visio 2016 & Visual Source Safe |
| **Operating System**: | Windows 10/8, Unix, Sun Solaris |
| **ETL/Data warehouse Tools:** | Informatica 9.6, SAP Business Objects XIR3.1/XIR2, Talend, Tableau, Ataccama |
| **Methodologies:** | RAD, JAD, RUP, UML, System Development Life Cycle (SDLC), Agile, Waterfall Model. |

# Client: Bank of America, Chicago IL

# Duration: July 2024 - Till Date

**Role: Sr Azure Data Engineer**

# Responsibilities:

* Involved in building database Models and Views utilizing **Python**, to build an interactive web-based solution.
* Collaborated with other developers to handle complicated issues related to the deployment of Django-based applications.
* Handled development and management of front-end user interfaces with the help of **HTML**, **CSS**, **jQuery,** and **JavaScript**.
* Modify the existing **Python/Django** modules to deliver certain formats of data and add new features.
* Automated a reporting process, using **Python**, **Luigi** (library for task workﬂow and dependencies), and other APIs.
* Written **Python** scripts using **Python** libraries such as **pandas**, and NumPy that do read/write operations on large CSV ﬁles, perform data aggregations and compare data by columns.
* Developed real-time data processing solutions using Databricks Streaming (Structured Streaming), processing streaming data from Kafka, Kinesis, or other message brokers.
* Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) Big Query.
* Experience in integrating **Python REST API** Framework using **Django**.
* Working experience in **Data Warehouse ETL /design** and implementation of complex big data pipelines.
* Used **Python**, **PySpark**, shell script, **oracle** scheduler, **Luigi**, **Oracle Pl SQL,** etc.
* Developed JSON scripts for deploying pipelines in Azure Data Factory (ADF) that process data using Azure Synapse SQL Activity, demonstrating experience with data pipeline development and integration with Azure services.
* Designed and optimized streaming workflows for low-latency data ingestion, transformation, and analytics in Databricks notebooks.
* Designed and designed ETL processes transforming and loading complicated data structures using AbInitio GDE.
* Collaborated extensively with payment product teams to incorporate new features and transaction workflows into the data architecture.
* Used Azure Databricks for processing large-scale data using distributed computing, including data ingestion, data transformation, and data analysis tasks, showcasing proﬁciency in utilizing **Azure** **Databricks** for big data processing
* Developed and implemented machine learning models using **Azure** **Databricks**, leveraging its built-in machine learning libraries and distributed computing capabilities, demonstrating expertise in machine learning on the **Azure** platform.
* Involved in migration of an Oracle SQL ETL to run on the Google Cloud platform using cloud Dataproc & Big Query, cloud pub/sub for triggering the Apache Airflow jobs
* Design, develop, and manage the Data Warehouse in Redshift, Snowflake & Data Lake for Analytics and Reporting
* As a Data Engineer I am responsible for building scalable distributed data solutions using Hadoop.
* Leveraged **Azure** **Databricks** notebooks for interactive data exploration, visualization, and analysis, showcasing proﬁciency in utilizing the collaborative notebook environment for data analysis and exploration.
* Integrated **Azure** **Databricks** with other **Azure** services, such as **Azure** Synapse Analytics, **Azure** Blob storage, and **Azure** **SQL** Database, for seamless data processing and analysis workﬂows, demonstrating proﬁciency in building end-to-end data solutions using **Azure** **Databricks**.
* A thorough understanding of financial data structures and payment lifecycle procedures, such as chargebacks, refunds, and settlements.
* Building Data pipelines & Data integration using Snowflake’s snow pipes.
* Utilized **Azure** **Databricks** for real-time data processing and stream analytics, showcasing expertise in processing and analyzing data in real-time using Azure Databricks streaming capabilities.
* Designed workflows using Airflow to automate the services developed for Change data capture.
* Implemented data quality checks and validation routines within Databricks pipelines to ensure data accuracy, completeness, and consistency.
* Familiar with using API integration to collect payment data from third-party platforms and ensure data consistency across systems.
* Proficient in designing and developing ETL jobs using IBM DataStage to extract, transform, and load data from heterogeneous sources (DB2, Oracle, SQL Server).
* Implemented monitoring and alerting solutions for Databricks workloads using tools like Prometheus, Grafana, or Databricks Monitoring, to proactively identify and address performance issues.
* Used Jenkins to deploy our code into diﬀerent environments and scheduling jobs.
* Expertise in Apache Spark and Databricks, including Spark SQL, Data Frame API, and Spark MLlib, for distributed data processing, analytics, and machine learning.
* Used bug-tracking tools like Jira, conﬂuence, and version controls Git, GitLab.

**Environment**: **Python**, Django, Luigi, windows, Linux, My**SQL**, **SQL**, Cassandra, AWS RDS, AWS S3, AWS EC2, Ka a, JSON, Restful API, MVC architecture, GitLab, Agile, PostgreSQL, Enterprise Scheduler, Bitvise SSH Client, Scrum, JIRA, GIT.

# Client: HCSC, Chicago IL

# Duration: Jan 2024 – June 2024

**Role: Data Engineer**

# Responsibilities:

* As a Data Engineer I am responsible for building scalable distributed data solutions using Hadoop.
* Involved in the Agile Development process (Scrum and Sprint planning).
* Handled Hadoop cluster installations in Windows environment.
* Migrated on-premises environment in **GCP** (Google Cloud Platform)
* Migrated data warehouses to Snowﬂake Data warehouse.
* Deﬁned virtual warehouse sizing for Snowﬂake for diﬀerent types of workloads.
* Integrated and automated data workloads to Snowﬂake Warehouse.
* Extensive experience in building and maintaining data pipelines on AWS Databricks using **Python** and SQL.
* Utilized Dbt Cloud/Core to architect and develop data models, ensuring scalability, flexibility, and maintainability of data pipelines for analytics.
* Process batch and real-time data using AbInitio components including Graphs, Plans, and Continuous Flows.
* Proficient in **Snowflake's** data loading mechanisms, including bulk loading and Snow pipe.
* As a Data Engineer I am responsible for building scalable distributed data solutions using Hadoop.
* Proficiently managed resources in Spark-on-Kubernetes by leveraging the improvements brought by the Kubernetes Resource Staging Server (RSS), ensuring efficient resource allocation and management
* Hands-on experience building **PySpark**, Spark Java, and Scala applications for batch and stream
* Created tables in snowﬂake DB, loading and analyzing data using Spark-Scala scripts.
* Developed ETL pipelines in and out of the data warehouse using a combination of **Python** and Snowﬂake’s Snow **SQL**.
* Written POCs in **Python** to analyze the data quickly before applying big data solutions to process at a scale.
* Responsible for data governance rules and standards to maintain the consistency of the business element names in th
* Build a Data Warehouse in the **Azure** platform using **Azure** data bricks and data factory.
* Developed data pipeline using Sqoop to ingest cargo data and customer histories into HDFS for analysis.
* Designed ETL using Internal/External tables and stored in parquet format for eﬃciency.
* Involved in porting the existing on-premises Hive code migration to **GCP** (Google Cloud Platform) Big Query.
* Involved in migration of an Oracle **SQL** ETL to run on the Google cloud platform using cloud Dataproc & Big Query, cloud pub/sub for triggering the Apache Airﬂow jobs.
* Experience in designing and implementing **Snowflake** data models, schemas, tables, and views to support efficient data organization and querying.
* Extracted data from data lakes, EDW to relational databases for analyzing and getting more meaningful insights using
* **SQL** Queries and PySpark.
* Experience with real-time streaming data processing using **AWS Databricks** Streaming and integrating with AWS services like Kinesis or **Kafka.**
* Later Migrated applications from Django to Flask and NoSQL (DynamoDB) to SQL**(Snowflake).**
* Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop
* cluster on AWS EMR.
* Designed, developed, and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems.
* Developed MapReduce programs to parse the raw data, populate staging tables, and store the reﬁned data in partitioned tables in the EDW.
* designed and implemented Data Vault 2.0 methodologies using the automated package within Dbt, ensuring scalability and flexibility in data warehousing solutions.
* Wrote Sqoop Scripts for importing and exporting data from RDBMS to HDFS.
* Set up Data Lake in Google Cloud using Google Cloud Storage, Big Query, and Big Table.
* Developed scripts in Big Query and connected them to reporting tools.
* Designed workﬂows using Airﬂow to automate the services developed for Change data capture.
* Carried out data transformation and cleansing using **SQL** queries and PySpark.
* Used Ka a and Spark streaming to ingest real-time or near real-time data in HDFS.
* Worked related to downloading Big Query data into Spark data frames for advanced ETL capabilities.
* Built reports for monitoring data loads into **GCP** and drive reliability at the site level.
* Participated in daily stand-ups, bi-weekly scrums, and PI panning.

**Environment:** Hadoop, **GCP**, Big Query, Snowflake DB, Big Table, Spark, Sqoop, ETL, HDFS, Snowﬂake DW, Oracle **SQL**, MapReduce and Agile process.

**Client:** **Abbott Laboratories, Green Oaks, IL**

**Duration: March 2023 – Dec 2023**

**Role: Big Data Engineer**

**Responsibilities:**

* Worked in an **Agile** environment, and used the rally tool to maintain the user stories and tasks.
* Implemented **Apache Sentry** to restrict access to the hive tables on a group level.
* Designed and implemented by configuring Topics in the new **Kafka** cluster in all environments.
* Created multiple dashboards in **Tableau** for multiple business needs.
* Implemented Partitioning, Dynamic Partitions, and Buckets in **HIVE** for efficient data access.
* Designed **SSIS Packages** to extract, transfer, and load (**ETL**) existing data into **SQL Server** from different environments for the **SSAS cubes (OLAP)**
* Extract Transform and Load data from source systems to **Azure Data Storage** services using a combination of **Azure Data Factory**, **T-SQL**, **Spark SQL**, and **Azure Data Lake** analytics.
* Designed & implemented database solutions in **Azure SQL Data Warehouse**, **Azure SQL**.
* Implemented Composite server for the data virtualization needs and created multiples views for restricted data access using a **REST API**.
* Vertical and horizontal scaling strategies were used to optimize the use of resources on **GKE clusters**, effectively controlling infrastructure costs while maintaining performance.
* Exported the analyzed data to the relational databases using **Sqoop** for visualization and to generate reports for the BI team Using **Tableau**.
* Migrated **Map reduce** jobs to **Spark jobs** to achieve better performance.
* Involved in converting **Map Reduce** programs into **Spark** transformations using **Spark RDDs** using **Scala and Python.**
* Worked cross-functionally with software engineers and data scientists to design customized **GKE** solutions that satisfied particular data processing specifications, ensuring peak performance.
* Developed **Apache Spark** applications by using **Spark** for data processing from various streaming sources.
* Developed data pipeline using **Spark**, **Hive**, **Pig**, **Python**, **Impala**, and **HBase** to ingest customer
* Involved in converting **Hive/SQL** queries into **Spark** transformations using **Spark RDDs**, **Python,** and **Scala**.
* Queried and analyzed data from **Cassandra** for quick searching, sorting, and grouping through **CQL**.
* Joined various tables in **Cassandra** using **Spark** and **Scala** and ran analytics on top of them.
* **GKE** configurations, deployment processes, and troubleshooting procedures were kept in-depth documented.
* Applied **Spark** advanced procedures like text analytics and processing using the in-memory processing.
* Implemented Apache Drill on **Hadoop** to join data from SQL and No SQL databases and store it in **Hadoop**.
* Brought data from various sources into Hadoop and Cassandra using **Kafka**.
* **SQL Server reporting services** (**SSRS**). Created & formatted Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Sub reports, ad-hoc reports, parameterized reports, interactive reports & custom reports.
* Created **Kubernetes** Deployments and Services to manage and expose containers, ensuring high availability and load balancing.
* Designing and Developing **Oracle PL/SQL** and **Shell Scripts**, Data Import/Export, Data Conversions and Data Cleansing.

**Environment:** **Hadoop, Map Reduce, HDFS, Hive, Pig, Impala, Kafka, Cassandra, Spark, Scala, Python, SQL, Zookeeper, Sqoop, Azure.**

# Client: Sunera Technologies, India

# Duration: June 2018 - Jul 2022

**Role: Data Engineer**

# Responsibilities:

* Worked as Data Engineer to review business requirements and compose sources to target data mapping documents.
* Conducted technical orientation sessions using documentation and training materials.
* Gathered the business requirements from the Business Partners and Subject Matter Experts.
* Served as technical expert guiding choices to implement analytical and reporting solutions for clients.
* Worked closely with the business, other architecture team members, and global project teams to understand, document and design data warehouse processes and needs.
* Implemented Installation and conﬁguration of multi-node cluster on Cloud using Amazon Web Services (AWS) on EC2.
* Developing code using Apache Spark and Scala, IntelliJ, NoSQL databases (Cassandra), Jenkins, Docker pipelines, GITHUB, Kubernetes, HDFS file System, Hive, Kafka for streaming Real-time streaming data, Kibana for monitor logs, etc. authentication/authorization to the data Responsible to deployments to DEV, QA, PRE-PROD (CERT) and PROD using AWS.
* Developed reconciliation process to make sure elastic search index document count matches source records.
* Maintained Tableau functional reports based on user requirements.
* Created action ﬁlters, parameters, and calculated sets for preparing dashboards and worksheets in Tableau.
* Used Agile (SCRUM) methodologies for Software Development.
* Regular audits and assessments were conducted to ensure that data assets were in compliance with existing data governance frameworks.
* Hasura now uses remote schemas to interface with existing APIs, allowing for a consistent data graph across microservices.
* Developed and implemented data governance rules, standards, and procedures to enhance overall data management and stewardship.
* Developed data pipelines to consume data from Enterprise Data Lake (MapR Hadoop distribution - Hive tables/HDFS) for analytics solution.
* Developed and implemented data governance rules, standards, and procedures to enhance overall data management and stewardship.
* Created Hive External tables to stage data and then move the data from Staging to main tables.
* Wrote complex Hive queries to extract data from heterogeneous sources (Data Lake) and persist the data into HDFS.
* Implemented the Big Data solution using Hadoop, hive, and Informatica to pull/load the data into the HDFS system.
* Developed incremental and complete load **Python** processes to ingest data into Elastic Search from Hive.
* Pulled the data from the data lake (HDFS) and massaged the data with various RDD transformations.
* Created Oozie workﬂow and Coordinator jobs to kick oﬀ the jobs on time for data availability.
* Developed Rest services to write data into Elastic Search index using **Python** Flask speciﬁcations
* Developed complete end-to-end Big-data processing in the Hadoop eco system.
* Used AWS Cloud with Infrastructure Provisioning / Conﬁguration.
* Created dashboards for analyzing POS data using Tableau.
* Developed Tableau visualizations and dashboards using Tableau Desktop.
* Involved in PL/**SQL** query optimization to reduce the overall run time of stored procedures.
* Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting on the dashboard.
* Continuously tuned Hive UDFs for faster queries by employing partitioning and bucketing.
* Implemented partitioning, dynamic partitions, and buckets in Hive.
* Deployed RMAN to automate backup and maintain scripts in the recovery catalog.
* Worked on QA the data and adding Data sources, snapshots, and caching to the report.

**Environment:** AWS, **Python**, Agile, Hive, Oracle 12c, Scala2.1.1, Tableau, HDFS, PL/**SQL**, Snowﬂake, Sqoop, Flume

# Client: DBS Bank, India

# Duration: Aug 2013 – May 2018

# Role: Data Analyst

# Responsibilities:

* Eﬀectively led client projects. These projects contained heavy **Python**, **SQL**, Tableau, and data modeling.
* Performed data merging, cleaning, and quality control procedures by programming data object rules into a database management system.
* Created detailed reports for management.
* Reported daily on returned survey data and thoroughly communicated survey progress statistics, data issues, and their resolution.
* Involved in Data analysis and quality check.
* Extracted data from source ﬁles, transformed and loaded to generate CSV data ﬁles with **Python** programming and **SQL** queries.
* Stored and retrieved data from data warehouses.
* Created the source-to-target mapping spreadsheet detailing the source, target data structure, and transformation rule around it.
* Trained and coached business users and analysts on data stewardship best practices and technologies, instilling a sense of data ownership.
* Designed and designed an end-to-end ETL framework utilizing IBM DataStage to cut data processing time by 60% while maintaining excellent data quality across numerous sources.
* Led the conversion of legacy ETL operations to IBM DataStage, which increased data integration efficiency and cut maintenance overhead by 40%.
* Worked on writing Scala programs using Spark on Yarn for analyzing data.
* Wrote **Python** scripts to parse ﬁles and load the data in a database, used **Python** to extract weekly information from the ﬁles, and Developed **Python** scripts to clean the raw data.
* Worked extensively with the Tableau Business Intelligence tool to develop various dashboards.
* Worked on datasets of various ﬁle types including HTML, Excel, PDF, Word, and its conversions.
* Analyzed data from company databases to drive optimization and improvement of product development, marketing techniques, and business strategies
* Developed Spark Streaming Jobs in Scala to consume data from Kafka topics, made transformations on data, and inserted to HBase.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
* Performed Database and ETL development per new requirements as well as actively involved in improving overall system performance by optimizing slow-running/resource-intensive queries.
* Developed data mapping documentation to establish relationships between source and target tables including transformation processes using **SQL**.
* Participated in data modeling discussion and provided inputs on both logical and physical data modeling.
* Reviewed the Performance Test results to ensure all the test results meet requirement needs.
* Created a master Data workbook that represents the ETL requirements such as mapping rules, physical Data element structure, and their description.

**Environment**: Oracle 10g, UNIX Shell Scripts, MS Excel, Scala, MS PowerPoint, **Python**, **SQL**.