**Vinay Krishna**

**4692155082**

**Vinaykrishna531@gmail.com**

[**https://www.linkedin.com/in/vinay-k-651b59276/**](https://www.linkedin.com/in/vinay-k-651b59276/)

**Summary:**

* **10+ years** of working experience in **Data Engineering, Data Pipeline Design, Development,** and **Implementation**.
* Strong experience in writing scripts using **Python API, PySpark API** and **Spark API** for analysing the data.
* **7+ years** of Experience in using various **Azure services**, such as **Data Factory, Data Bricks**,**Azure Blob Storage, Azure Synapse Analytics,** Pub/Sub, Dataflow, **Azure Virtual Machines**, etc.
* **8 years** of Experience in using various **AWS services**, such as **Redshift** , **EC2 ,** S3, AWS Glue.
* Experience in building data pipelines using **Azure Data Factory, Azure Databricks**, and loading data to **Azure Data Lake, Azure SQL Database, Azure SQL Data Warehouse** to control and grant database access.
* Experience with both **AWS, AWS Glue & Azure** (**Data Factory, Databricks & Data Lake) cloud**.
* Good experience with **Azure** services like **HDInsight**, **Stream Analytics**, **Active Directory**, **Blob** **Storage, Cosmos DB, Storage Explorer**.
* Developed **Spark** Applications that can handle data from various **RDBMS (MySQL, Oracle Database)** and Streaming sources.
* Experienced in developing **Web-based Applications** using **Python, CSS, HTML, JavaScript**.
* Implemented different **Python** libraries **Pandas**, **NumPy**, **matplotlib**, **python-twitter**, **Panda’s** **data frame** in various tasks.
* Hands On experience on **Spark Core, Spark SQL, Spark Streaming** and creating the **Data Frames** handle in **SPARK** with **snowflake**.
* Expertise in developing multiple confluent **Kafka** Producers and Consumers to meet business requirements. Store the stream data to **HDFS** and process it using **Spark**.
* Experienced working on Hadoop/ Bigdata Framework and its ecosystem like HDFS, MapReduce, Yarn,Spark, Hive, impala, Sqoop and Oozie.
* Strong Hadoop / Bigdata and platform support experience with all the entire suite of tools and services in major **Hadoop Distributions** – **Cloudera**, **Amazon EMR**, **Azure HDInsight**, and **Hortonworks**.
* Hands-on use of **Spark** and **Scala API’s** to compare the performance of **Spark** with **Hive** and **SQL**, and **Spark SQL** to manipulate **Data Frames** in **Scala**.
* Experienced in scripting with **Python (PySpark), Scala** and **Spark-SQL** for development, aggregation from various file formats such as **XML, JSON, CSV, Parquet**.
* Expertise in writing **Hadoop** Jobs using **MapReduce**, **Apache** **Crunch**, **Hive**, **Pig**, and **Splunk**
* Experience in developing Data pipelines using **AWS** services including **EC2, S3, Redshift, Glue, Lambda functions, Step functions, CloudWatch, SNS.**
* Expertise in migrate an existing on-premises application to **AWS**. Used **AWS** **services** like **EC2** and **S3** for small data sets processing and storage
* Proficient with **Spark Core**, **Spark SQL** for processing and transforming complex data using in-memory computing capabilities written in **Python**.
* Strong understanding of **AWS components** such as **EC2 and S3**.
* programming experience with **Scala, Java, Python, SQL, Spark** Developed database objects like **Stored Procedures** and **Packages** using **SQL**.
* Used **PySpark** and Pandas to load Data from various sources & Various structures
* Experience in building data pipelines using **Azure Data Factory, Azure Databricks**, and loading data to **Azure Data Lake, Azure SQL Database, Azure SQL Data Warehouse** to control and grant database access.
* Expertise in different **AWS services, CI/CD** implementations and tools
* Expertise in migrate an existing on-premises application to **AWS**. Used **AWS** services like **EC2 and S3** for small data sets processing and storage
* Experienced with **Spark** to improve efficiency of existing algorithms using **Spark Context, Spark SQL, Data Frame, Pair RDD's** and **Spark YARN.**
* Experienced with databases, **SQL queries**, establishing data relationship and performance tuning for data extract.
* Expertise in use of Sqoop to migrate data between **RDBMS**, **NoSQL databases** and **HDFS**
* Experience in devops tools like **Jenkins, Selenium**
* Experience in **Apache Spark, SparkSQL, Spark DataFrames, Spark RDD’s**.
* Responsible for managing data from multiple sources
* Expertise in creating **Kubernetes** cluster with cloud formation templates and **PowerShell** scripting to automate deployment in a cloud environment.
* Hands-on experience on **Ad-hoc queries**, **Indexing, Replication, Load balancing, Aggregation** in **MongoDB**
* Experience in using **bug tracking** and ticketing systems such as **Jira, used Git for version control.**
* Involved in migration of the legacy applications to cloud platform using DevOps tools like **GitHub, Jenkins, JIRA and Docker.**
* Experience in using **Docker** and **Ansible** to fully automate the deployment and execution of the benchmark suite on a cluster of machines
* Experience in various methodologies like Waterfall and Agile.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming & Scripting** | Python, Scala, Java, SAS, R, SQL, MATLAB, HiveQL, PowerShell, and BASH Scripting |
| **Python Libraries** | Pandas, SciPy, NumPy, Scikit-Learn, Stats Models, Matplotlib, Plotly, Seaborn, TensorFlow, PyTorch, PySpark |
| **Query languages** | SQL, T-SQL/SQL, PL/SQL, HiveQL, GraphQL |
| **Big Data Ecosystem** | HDFS, YARN, MapReduce, Sqoop, Hive, Oozie, Pig, Spark, Zookeeper, Cloudera Manager, Kafka, Flume, Nifi, Connect, Airflow, Stream Sets, Kafka connect, confluent |
| **AWS** | EMR, EC2, EBS, RDS, S3, Athena, Glue, Elasticsearch, Lambda, SQS, DynamoDB, Redshift, PostgreSQL |
| **Azure** | Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Azure Active Directory, Databricks. |
| **ETL Tools** | Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, AtScale, Power BI, Arcadia |
| **Data warehouses** | Snowflake, Amazon Redshift, Google BigQuery, Microsoft Azure Synapse Analytics, Oracle Autonomous Data Warehouse, IBM Db2 Warehouse, Teradata, SAP BW/4HANA, Informatica PowerCenter, Apache Hive. |
| **SQL Databases and ORM** | Oracle, MySQL, Teradata, Postgres, Django ORM, SQL Alchemy |
| **NoSQL Database** | HBase, Cassandra, PostgreSQL Dynamo DB, Cosmos DB, Mongo DB |
| **CI/CD/Build Tools** | Jenkins, Maven, Ant |
| **Version Control** | Git, SVN |
| **BI Reporting Tools** | Tableau, Power BI, Looker |
| **Operating Systems** | Linux, Unix, Windows 10, Windows 8, Windows 7, Windows Server 2008/2003, Mac OS |

**Certifications:**

* **Azure Data Engineer**
* **SnowPro Core Certified**

**Work Experience:**

**Client: 7-11, Scottsdale, AZ Jul 2021 – till date**

**Role: Azure Data Engineer**

**Responsibilities:**

* We got requirement to migrate complete **data** from **MySQL, SQL Server** and from **APIs** to **cloud database Snowflake** using **cloud Azure**.
* For which we have developed ETL pipelines according to the Business Requirement in **Azure Data Bricks** (**ETL**).
* Developed **Data ingestion** modules (both real time and batch data load) to data into various layers in **GCS, BigQuery** and **Snowflake** using **Google Cloud Data Flow, and Work Flow.**
* Proficient in Devops Operation.
* This project has given me a tremendous opportunity by introducing to **Cloud**. In this, I have acquired the knowledge starting from requirement gathering to the designing the job and scheduling it.
* Experienced in leveraging **Apache Spark** capabilities within **Azure Databricks** for distributed data processing tasks, including data shuffling and partitioning.
* I also have the functional knowledge of the retail domain and have done part as analyst in cases.
* We have developed daily jobs which gets file from **Azure Blob storage** instead of regular databases and loaded it into **Azure cloud database** with the required logics.
* Developed **Spark** Applications that can handle data from various **RDBMS (MySQL, Oracle Database)** and Streaming sources.
* Developed **Data ingestion** modules (both real time and batch data load) to data into various layers in **Azure Blob Storage, Azure Synapse Analytics** and **Snowflake** using **Azure Data Factory, and Azure Logic Apps**
* Loaded **Data** from **API** to **Azure Blob Storage** and then copied into **Snowflake,** used **Python** for the same as a programming language.
* **Data Extraction, aggregations**, and consolidation of **Data** within **Azure Data Factory** using **PySpark**.
* Developed a **Python** script to transfer Data from on-premises to **Azure Blob Storage.**
* Developed a **Python** script to hit **REST API's** and extract data to **Azure Blob Storage.**
* Developed **Python** scripts to update content in the database and manipulate files.
* Developed jobs to load data from **Azure Blob Storage /Snowflake** to servers and excel to **Snowflake/Blob Storage.**
* Developed **Python** code to load semi-structured to **Snowflake**, from **JSON**, **XML** etc.
* Created tasks and streams.
* Created reader accounts for different sharing to specific business needed roles.
* Created **Snowpipes** to ingest Data as they get loaded into **Blob Storage** using **SNS Queue**.
* Migrated jobs from **SQL Server** to **Cloud** jobs using the same logic.

**Environment:** Azure, Snowflake, Mysql, azure Blob Storage , Sqlserver, Azure Synapse Analytics, Azure Data Factory, Json, XML, Spark, Pyspark, Python, Anaconda Navigator, Tableau, Putty.

**Client: Amgen, Santa Clara, CA Aug 2020 to June 2021**

**Role: Data Engineer**

**Responsibilities:**

* We got requirement to Migrate complete **Data** from **Oracle, MySQL, SQL Server** and from **APIs to Snowflake** hosted in AWS.
* We created Storage and containers specific to the applications/business needs and created integration connection between **DB** and **AWS**.
* Using JDBC drivers integrated snowpipe in snowflake and Tableau and created charts for real time analytics to report senior executives.
* Created interactive reports using powerBI with data consisting of million records and 1000 entities of historical pharma data which improved data efficiency by 20%.
* Built ETL pipelines in AWS Glue for transferring 5 million records from multiple sources and transfred that data into **AWS S3.**
* Led data mapping initiatives to ensure seamless data flow between different systems, enhancing data accuracy and consistency.
* Collaborated with IT teams to identify data sources and define mapping rules for ETL (Extract, Transform, Load) processes.
* Performed day-to-day integration with the **Database Administrators, SQL Server**, **Oracle** and **Azure Cloud** teams to ensure the insertion of **database tables**, columns and its **metadata** have been successfully implemented out to the **DEV, QUAL** and **PROD** region environments in **AWS Cloud – Snowflake**.
* Created work spaces and built ETL Pipelines in **AWS Glue** which migrated 50 Terabytes of data from SQL Server to **AWS Redshift.**
* Experience in Developing **Spark** applications using **Spark - SQL** in **Databricks** for **data extraction**, **transformation**, and **aggregation** from multiple file formats for analysing & transforming the data to uncover insights into the customer usage patterns.
* We have developed jobs according to the requirement from the client using (ETL).
* Created snowpipe to load continuous Data to DB as early as they get loaded into the container.
* I also have the functional knowledge of the banking domain and have done part as analyst in cases.
* We have developed daily jobs which gets file from container instead of regular databases and loaded it into cloud database with the required logics.
* We have developed jobs for incremental (with the help of audit table) as well as history (for large tables using iterator).
* Loaded **Data** from **API** to container and then copied into **Snowflake**, used **Python** for the same as a programming language.
* Developed jobs to load **Data** from **Container /Snowflake** to **servers** and excel to **Snowflake** **Container**.
* Developed **Python** code to load semi-structured to **Snowflake**, from **JSON, XML** etc. by creating stage and specific file formats.

**Environment:** AWS,Snowflake, Databricks, spark, Snowflake, Redshift ,AWS Glue, Mysql, Sqlserver, Json, XML, Python, Python, Tableau, powerbi,Snaplogic Anaconda Navigator, Tableau, Putty.

**Client: Novo, San Mateo, CA Apr 2019 – Aug 2020**

**Role: Azure Data Engineer**

**Responsibilities**:

* Collaborated with **Business Analysts, SMEs** across departments to gather business requirements, and identify workable items for further development.
* Created **Azure data factory** instances, including defining linked services, datasets and integration runtimes for seamless data orchestration.
* We created Storage and containers specific to the applications/business needs and created integration connection between **DB** and **GCP**.
* Created work spaces and built ETL Pipelines in **AWS Glue** which migrated 50 Terabytes of data from SQL Server to **AWS Redshift.**
* Skilled at developing and maintaining **Azure Databricks** notebooks using programming languages like Python, Scala, or R for data analysis, transformation, and model development.
* To build a web application by using **Django/Python** and while using **HTML/CSS/JS** for **server-side** rendered application.
* Created a **Python/Django**-based **web application** using **Python** scripting for data processing, **MySQL** Workbench for the **database**, and **HTML/CSS** and High Charts for data visualization of the served pages.
* Developing **Python** Framework using **Django** to perform scan software unit monitoring.
* Designed the front end of the application using **Python**.
* Implemented **React JS components**, Forms, Events, Keys
* Partnered with **ETL developers** to ensure that Data is well cleaned, and the **Data warehouse** is up to date.
* Rewrite existing **Python** modules to deliver certain formats of data. Created script in **Python** for calling **REST APIs**.
* Responsible for setting up **Python REST API** **framework** using **Django**.
* Selected and generated data into **CSV files** and stored them into **Azure Blob Storage**  by using **Azure Virtual Machines** and then structured and stored in **Azure Synapse Analytics**.
* Used **PySpark** and **Pandas** to calculate the moving average and **RSI** score of the stocks and generated them into **data warehouse**.
* Exploring with **Spark** to improve the performance and optimization in **Hadoop** using **Spark** context, **Spark-SQL, postgreSQL, Data Frame, pair RDD's**
* Involved in integration of **Hadoop cluster** with spark engine to perform BATCH operations.
* Performed **data pre-processing** and feature engineering for further predictive analytics using **Python Pandas**.

**Environment:** Spark (PySpark, SparkSQL,), Hadoop, Python 3.x (Numpy, Pandas), Tableau 10.1, GitHub, Azure Data Factory, Azure Databricks, spark, Snowflake, Teradata,

**Client:Tronox, San Diego, CA Jan 2013 – Mar 2019**

**Role: Data Engineer**

**Responsibilities:**

* Involved in Different projects which include two cloud environments GCP and Azure from **HDFS, Hive** and Third party **APIs** to **cloud database GCP and Azure**.
* Analysed large and critical datasets using **HDFS**, **HBase**, **MapReduce**, **Hive**, **Hive UDF**, **Pig, Sqoop, Zookeeper and Spark**.
* Selected and generated data into **CSV files** and stored them into **Google Cloud Storage** by using **Google Compute Engine** and then structured and stored in **BigQuery**.
* Utilized **Azure data factory** data flow activities and transformations to cleanse, reshape and enrich data during ETL process.
* Loaded and transformed large sets of structured, semi structured, and **unstructured data** using **Hadoop/Big Data concepts**.
* Performed Data transformations in **HIVE** and used **partitions**, buckets for performance improvements.
* Developing **Spark scripts**, **UDF's** using both **Spark DSL** and **Spark SQL** query for **Data aggregation, querying**, and writing data back into **RDBMS** through **Sqoop**.
* Designed and developed a **Data Lake** using **Hadoop** for processing raw and processed claims via **Hive and Informatica**.
* Developed **Spark** code using **Scala** and **Spark-SQL/Streaming** for faster processing of **Data**.
* Ingested data into **HDFS** using **Sqoop** and scheduled an incremental load to **HDFS**.
* Using **Hive** to analyse data ingested into **HBase** by using **Hive-HBase** integration and compute various metrics for reporting on the dashboard.
* Worked with **Hadoop** infrastructure to storage data in **HDFS** **storage** and use **Spark HIVE** **SQL** to **migrate** underlying **SQL** codebase in **Azure**.
* Experience in testing **Big Data Hadoop (HDFS, Hive, Sqoop and Flume), Master Data Management (MDM) and Tableau Reports.**
* Wrote **Pig Scripts** to generate Map Reduce jobs and performed **ETL** procedures on the data in **HDFS**.
* Generate **metadata**, mappings to load **data warehouse, data lake.**
* Used Zookeeper to provide coordination services to the cluster.
* Analysed data using **Hive** the partitioned and bucketed data and compute various metrics for reporting.
* Built **Azure Data Warehouse** Table Data sets for **Power BI Reports**.
* Extract Transform and Load data from Sources Systems to **Azure Data Storage** services using a combination of **Azure, T-SQL, Spark SQL**, and **Azure Data Lake Analytics**
* **Data Ingestion** to one or more **Azure Services** - (A**zure Data Lake, Azure Storage, Azure SQL, Azure DW)** and processing the **Data in Azure Databricks**.
* Built **Cloud** **Data Lake** on **Snowflake**, **Redshift** and **Azure Sql Data warehouse** platforms with **Talend.**
* Working on **BI reporting** with At Scale **OLAP for Big Data.**
* Implemented **Kafka** for streaming data and filtered, processed the data.
* Designed and Developed Real time Stream processing Application using **Spark, Kafka, Scala**, and **Hive** to perform Streaming **ETL** and apply **Machine Learning**.
* Developed data pipeline using **flume, Sqoop** and pig to extract the **Data** from **weblogs** and store in **HDFS**.
* Developed **Shell scripts** for scheduling and Automating the job flow.
* Developed a workflow using Nifi to automate the tasks of loading the **data into HDFS**.
* Developed Map Reduce jobs to calculate the total usage of data by commercial routers in different locations, developed Map reduce programs for data sorting in **HDFS**.
* Load balancing of **ETL processes**, database performance tuning ETL processing tools.
* Loaded the data from **Teradata** to **HDFS** using **Teradata** **Hadoop** connectors.

**Environment:** Spark, HIVE, HDFS, HBase, MapReduce, Sqoop, Flume, RDBMS, Informatica, Zookeeper, Pig, Scala, NiFi, Python, Hadoop, Teradata, ETL, Hadoop, Azure, NOSQL, Sqoop, MYSQL, Agile Scrum methodology, Azure, Snowflake, Redshift.