Anji Naik Jeshavath

Senior Data Engineer

+1 443-794-0141

Anjijeshavath1@gmail.com

Maryland, USA

Career Objective

As a highly skilled Data Engineer with extensive experience in leveraging Azure, AWS, and GCP tools, I aim to contribute data infrastructure by designing and implementing robust, scalable, and efficient data pipelines. Seeking to contribute my skills and knowledge to a forward-thinking organization, leveraging cloud technologies to drive data-driven decision-making and innovation.

Profile Summary

- **5+** years of expertise designing, developing, and executing data pipelines and data **lake** requirements in numerous companies using the Big Data Technology stack, **Python**, **PL/SQL**, **SQL**, REST **APIs**, and the **Azure** cloud platform.
- Experience in Evaluation/design/development/deployment of additional technologies and automation for managed services on **S3**, **Lambda**, **Athena**, **EMR**, Kinesis, SQS, SNS, CloudWatch, Data Pipeline, **Redshift**, Dynamo DB, **AWS Glue**, Aurora DB, **RDS**, **EC2**. Experience in working with RIVERY **ELT** platform which performs data integration, data orchestration, data cleansing, and other vital data functionalities.
- Develop batch processing solutions by using **Data Factory** and **Azure Data bricks.**
- Working knowledge of **HDFS**, **Kafka**, Map Reduce, **Spark**, **PIG**, **HIVE**, **Sqoop**, **HBase**, **Flume**, and **Zookeeper** as tools for designing and deploying end-to-end big data ecosystems.
- Developed the infrastructure using the technologies like **Kafka**, **Splunk, CASSANDRA.**
- Develop batch processing solutions by using **Data Factory** and **Azure Data bricks.**
- Experienced with **JSON** based **RESTful web services**, and **XML/QML** based **SOAP** web services and also worked on various applications using **python** integrated **IDEs** like **Sublime Text** and **PyCharm**.
- Deep understanding of performance tuning, partitioning for optimizing **spark applications.** Experience in Performance Monitoring, Security, Trouble shooting, Backup, Disaster recovery, Maintenance and Support of **Linux** systems.
- Experience in **Agile** software development process, Test Driven Development and Scrum methodologies.
- Architect & implement medium to large scale **BI solutions** on **Azure** using **Azure** Data Platform services (**Azure Data Lake**, **Azure Data Factory**, **Data Lake Analytics**, and Stream Analytics).
- Proficient in building **CI/CD** pipelines in **Jenkins** using pipeline syntax and groovy libraries.
- Proficiency with **Scala**, **Apache HBase**, **Hive**, **Pig**, **Sqoop**, **Zookeeper**, **Spark**, **Spark SQL**, **Spark** Streaming, **Kinesis**, **Airflow**, Yarn, and **Hadoop** (**HDFS, MapReduce**).
- Experienced in building Snow Pipes, migrating **Teradata** objects into **Snowflake** environment.
- Used microservices and containerization technologies such as **Docker** and **Kubernetes** to build scalable and resilient SaaS applications. Experience working with Front end technologies like **Html**, **CSS**, **JS**, **ReactJS**.
- In depth knowledge about Data Warehousing (gathering requirements, design, development, implementation, testing, and documentation), Data Modelling (analysis using Star Schema and **Snowflake** for **FACT** and Dimensions Tables), Data Processing, Data Acquisition and Data Transformations (Mapping, Cleansing, Monitoring, Debugging, Performance Tuning and Troubleshooting **Hadoop clusters**).
- Hands-on experience in handling database issues and connections with **SQL** and **NoSQL** databases such as **MongoDB**, **HBase**, **Cassandra**, **SQL** server, and **PostgreSQL**. Ability to work effectively and efficiently as a team member as well as individually with a desire to learn new skills and technology.
- Experience in Cisco Cloud Center to more securely deploy and manage applications in multiple data center, private cloud, and public cloud environments. Worked on production support looking into logs, hot fixes and used Splunk for log monitoring along with **AWS** CloudWatch.

Education Details

University of Maryland Baltimore County, Maryland , USA

- Masters / Data Science (Aug 2022 - May 2024)

Technical Skills

- **Programming & Scripting:** Python, Java, Scala, SQL, Bash.
- **Big Data Technologies:** Hadoop, Apache Spark, Hive, HDFS, Presto, Pig, Flink
- **ETL/ELT Tools:** Informatica, Talend, Apache Nifi, dbt, SSIS
- **Data Warehousing:** Snowflake, Redshift, Azure Synapse, Big Query, Teradata, Vertica.
- **Data Integration & Orchestration:** Apache Airflow, Luigi, AWS Step Functions, Azure Logic Apps, GCP Data Fusion
- **Streaming & Real-Time Processing:** Apache Kafka, AWS Kinesis, Azure Event Hubs, Google Pub/Sub, Spark Streaming.
- **AWS:** Redshift, S3, Glue, Lambda, EMR, Athena, DynamoDB, Kinesis, CloudFormation
- **Azure:** Azure Data Factory, Synapse Analytics, Azure Storage, Databricks, Cosmos DB, Logic Apps, Azure Functions
- **Google Cloud (GCP):** Big Query, Dataflow, Pub/Sub, Cloud Storage, Cloud Composer, Cloud Functions
- **Database Systems:** PostgreSQL, MySQL, Oracle, SQL Server, MongoDB, Cassandra, DynamoDB, Cosmos DB, HBase
- **Business Intelligence Tools:** Tableau, Power BI, Looker, AWS Quick Sight, Google Data Studio
- **DevOps & CI/CD:** Jenkins, GitLab CI/CD, Azure DevOps, AWS Code Pipeline, Docker, Kubernetes
- **Machine Learning Integration:** AWS Sage Maker, Azure ML, GCP Vertex AI
- **Workflow Automation:** Terraform, CloudFormation, Ansible
- **Version Control & Collaboration:** Git, GitHub, GitLab, Bitbucket

Work History

Lennar Corporation - Irving, Texas, USA

Azure Data Engineer — Jul 2024 - Current

Lennar Corporation (Lennar) is a home construction company. I design and implement data warehouses using Azure Synapse Analytics (Dedicated SQL Pools), applying star and snowflake dimensional modelling techniques to support business intelligence and reporting needs. Perform schema design, indexing, and query optimization to ensure efficient access to large-scale analytical datasets.

Key Responsibilities:

- Performed **ETL** operations using **Python**, **Spark SQL**, **S3** and **Redshift** on terabytes of data to obtain customer insights.
- Experience in hive partitioning, bucketing and perform joins on hive tables and utilizing hive **SerDes** like **REGEX, JSON,** and **AVRO**. Involved in the automation process through **Jenkins** for **CI/CD pipelines**.
- Native integration with **Azure Active Directory** (Azure AD) and other **Azure** services enables to build modern data warehouse and machine learning and real-time analytics solutions.
- Led requirement gathering, business analysis, and technical design for **Hadoop** and **Big Data projects.**
- Develop **Spark** streaming application tread raw packet data from **Kafka** topics, format it to **JSON** and push back to **Kafka** for future use case's purpose. Created and provisioned different **Databricks** clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
- Worked with **Spark Core**, **Spark ML**, **Spark Streaming** and **Spark SQL** and **data bricks.**
- Created Data tables utilizing **PyQt** to display customer and policy information and add, delete, update customer records.
- Implemented a Reusable plug & play **Python** Pattern (Synapse Integration, Aggregations, Change Data Capture, Deduplication and High Watermark Implementation. This process accelerated the development time and standardization across teams. Developed and automated data migration pipelines using **Python**, **Apache Airflow**, and **GCP** services, ensuring data consistency and minimizing downtime during cutover.
- Involved in various phases of Software Development Lifecycle (**SDLC**) of the application, like gathering requirements, design, development, deployment, and analysis of the application. Written queries in **MySQL** and Native **SQL**.
- Built a common **SFTP** download or upload framework using **Azure Data Factory** and **Databricks**.
- Developed container-based **Docker**, worked with **Docker** images, **Docker** Hub and Docker-registries and **Kubernetes**.
- Documented and standardized **Elasticsearch** usage patterns, improving team onboarding speed and reducing query-related incidents by 40%.
- Implemented Performance tuning techniques in **Azure Data Factory** and **Azure Synapse Analytics**.
- Develop metrics based on **SAS** scripts on legacy system, migrating metrics to snowflake (**Google Cloud**).
- Implemented **RESTful** Web-Services for sending and receiving data between multiple systems.

Technologies Used: Airflow, Analytics, Apache, Azure, Azure Synapse Analytics, CI/CD, Data Factory, Docker, Elasticsearch, ETL, Factory, GCP, Jenkins, JS, Kafka, Kubernetes, lake, MySQL, Python, Redshift, S3, SAS, Services, Spark, Spark Core, Spark SQL, Spark Streaming, SQL

Hilton - Addison, Texas, USA
AWS Data Engineer **Sep 2022 - Jun 2024**

Hilton Hotels & Resorts is a well-known and respected brand in the hospitality industry, offering a wide range of full-service hotels and resorts worldwide. Leveraged EC2 and EMR for complex data transformations and processing when needed. Automated the data integration workflows using scripting languages like Python or Scala.

Key Responsibilities:

- Worked with **AWS** Terraform templates in maintaining the infrastructure as code.
- Strong knowledge of **ETL** best practices and experience designing and implementing **ETL** workflows using Talend.
- The **AWS Lambda** functions were written in **Spark** with cross - functional dependencies that generated custom libraries for delivering the **Lambda** function in the cloud. Performed raw data ingestion into, which triggered a lambda function and put refined data into **ADLS**. Hands on experience in data loading techniques using **Sqoop**.
- Used **PySpark**, **Apache Flink**, **Kafka**, and **Hive** on a distributed Hadoop **Cluster** to contribute to the creation of real- time streaming applications. Participated in development/implementing of **Cloudera Hadoop** development.
- Worked on **Kafka** streaming on subscriber side, processing the messages and inserting them into the db and **Apache Spark** for real-time data processing. Prepared scripts to automate the Ingestion process using **PySpark** and **Scala** as needed through various sources such as **API**, **AWS S3**, **Teradata** and **Redshift**.
- Involved in the entire lifecycle of the projects including Design, Development, and Deployment, Testing and Implementation, and support. Developed **T-SQL** scripts for managing instance-level objects and optimizing performance.
- Experience working with **Docker** to improve our (**CD**) Continuous Delivery framework to streamline releases
- Designed and executed data migration strategies for relational databases (**MySQL**, **PostgreSQL**) to cloud-native solutions like **BigQuery** and Cloud **SQL**, reducing query latency by 40%. Designed and deployed a **Kubernetes**-based containerized infrastructure for data processing and analytics, leading to a 20% increase in data processing capacity.
- Built robust data ingestion pipelines using **Logstash**, **Filebeat**, and **Kafka** Connect to stream real-time logs and events into **Elasticsearch** clusters. Utilized **Elasticsearch** and Kibana for indexing and visualizing the real-time analytics results, enabling stakeholders to gain actionable insights quickly.
- Proficient in using **Snowflake** utilities, **Snow SQL**, **Snow Pipe,** and applying Big Data modeling techniques using **Python**.
- Used **SSIS** to build automated multi-dimensional cubes.
- Stored the log files in **AWS S3**. Used versioning in **S3 buckets** where the highly sensitive information is stored.

Technologies Used: Apache, API, AWS, BigQuery, Cloudera, Cluster, Docker, Elasticsearch, ETL, Hive, Kafka, Kubernetes, lake, Lambda, MySQL, PostgreSQL, PySpark, Python, Redshift, S3, Scala, Snowflake, Spark, SQL, Sqoop, SSIS, Teradata

(Infosys Limited) Shutterfly - Bangalore, India
GCP Data Engineer **Aug 2021 - Aug 2022**

Shutterfly, LLC. is an American photography, photography products, and image sharing company. Integrated the data from various sources, including relational databases (like Cloud SQL), SaaS applications, and potentially on-premises systems, into the GCP data platform.

Key Responsibilities:

- Involved in loading data from **UNIX** file system to **HDFS**.
- Managed, Configured and scheduled resources across the cluster using **Azure Kubernetes** Service.
- Created reusable views and data marts in **BigQuery** to power Data Studio reports with consistent metrics and definitions.
- Experienced in working **Services** like Data **Lake**, Data **Lake Analytics**, **SQL** Database, Synapse, Data **Bricks**, Data factory, Logic Apps and **SQL** Data warehouse and **GCP** services Like Big Query, **Dataproc**, Pub sub etc.
- Employed Hadoop scripts to manipulate and load data from the **Hadoop File System.**
- Experienced in **GCP** features which include **Google Compute engine**, **Google Storage**, **VPC**, **Cloud Load balancing**, **IAM**.
- Processed the data efficiently with in **Azure Databricks** and visualizing insights through **Tableau** dashboards
- Experienced developing and deploying data solutions directly from **Cloud Shell environment**
- Developed **Spark** applications for the entire batch processing by using **PySpark**.

- Used **Python** based GUI components for the Front-End functionality such as selection criteria.
- Experienced in **Google Cloud components, Google container** builders and **GCP** client libraries and **Cloud SDK'S.**
- Involved in loading data into **Cassandra NoSQL** Database.
- Involved in monitoring and scheduling the pipelines using Triggers in **Azure Data Factory.**
- Migrating an entire oracle database to **BigQuery** and using of **Power BI** for reporting.
- Designed **Cassandra** schemas for time-series IoT data (500K writes/sec)
- Used **Sqoop** import/export to ingest raw data into Google Cloud Storage by spinning up **Cloud Dataproc cluster**.
- Used **Google Cloud Dataflow** using **Python sdk** for deploying streaming jobs in **GCP** as well as batch jobs for custom cleaning of text and json files and write them to **BigQuery**.
- Build data pipelines in **Airflow**/**Composer** for orchestrating **ETL** related jobs using different airflow operators.

Technologies Used: Airflow, Analytics, Azure, BigQuery, Bricks, Cassandra, ETL, Factory, GCP, HDFS, Kubernetes, Lake, Power BI, PySpark, Python, SDK, Services, Spark, SQL, Sqoop, Tableau, VPC

Bupa, Bangalore, India
Data Engineer **Sep 2019 - Jul 2021**

British United Provident Association Limited, is a British multinational health insurance and healthcare company. Worked independently to develop analytic applications leveraging technologies such as: Hadoop, NoSQL, In-memory Data Grids, Kafka, Spark, Ab Initio

Key Responsibilities:

- Designed **SSIS** control flow tasks for orchestrating the sequence and logic of **ETL** processes, such as conditional branching, looping, and error handling. Created **AWS Lambda** functions and assigned **IAM** roles to schedule **Python** scripts using CloudWatch Triggers to support the infrastructure needs (**SQS, Event Bridge, SNS**)
- Involved in creating **Hive** tables, loading, and analyzing data using hive scripts.
- Imported real time weblogs using **Kafka** as a messaging system and ingested the data to **Spark Streaming** and did data quality checks using **Spark Streaming** and arranged bad and passable flags on the data.
- Developed business logic using **Kafka** & **Spark Streaming** and implemented business transformations. Supported Continuous storage in ADLS and configured Snapshots and wrote entities in spark along with named queries to interact with database. Implemented DAGs for scheduling complex workflows, integrating with **AWS** services (S3, **Redshift**, Lambda) and external **APIs**.
- Developed **Databricks ETL** pipelines using notebooks, **Spark** Data frames, **Spark SQL** and **Python** scripting.
- Spark **SQL** to enable automated transformation of RDD case classes to schema RDD for both **Scala** and **Python** interfaces.
- Involved in requirement analysis, design, coding, and implementation phases of the project.
- Develop dashboards and visualizations to help business users analyse data as well as providing data insight to upper management with a focus on **Microsoft** products like **SQL** Server Reporting **Services** (SSRS) and Power BI.
- Instantiated, created, and maintained **CI/CD** continuous integration & deployment pipelines and apply automation to environments and applications. Analyzed data using Hadoop components **Hive** and **Pig**.
- Used **Apache** airflow in **GCP** composer environment to build data pipelines and used various airflow operators like bash operators, Hadoop operators and python callable and branching operators.
- Developed automated monitoring and alerting systems using **Kubernetes** and **Docker**, ensuring proactive identification and resolution of data pipeline issues. Optimized **Elasticsearch** cluster performance through shard tuning, heap memory management, refresh interval adjustments, and query profiling.
- Provisioned high availability of **AWS EC2** instances, migrated legacy systems to **AWS**, and developed Terraform plugins, modules, and templates for automating **AWS** infrastructure. Developed **ETL** pipelines between data warehouses using a combination of **Python** and **Snowflake**, **SnowSQL**, writing **SQL** queries against **Snowflake**.

Technologies Used: Apache, API, AWS, CI/CD, Cosmos DB, , Docker, EC2, Elasticsearch, ETL, Gateway, GCP, HDInsight, Hive, IaaS, Kafka, Kubernetes, Lambda, Microsoft, PaaS, Pig, Power BI, Python, Redshift, S3, Scala, Services, Snowflake, Spark