Shiva Reddy

9373150760 **Data Engineer** **nareddyshivareddy06@gmail.com**

PROFESSIONAL SUMMARY

- 5+ years of experience in IT, which includes experience in Bigdata Technologies, Hadoop ecosystem, Data Warehousing, SQL related technologies in various sectors. 3 Years of experience in Big Data Analytics using Various Hadoop eco-systems tools and Spark Framework and Azure cloud services using Scala and python as the main programming dialect. 4 years of experiences on Data warehouse developer role.
- Proficient in Python programming with hands-on experience in implementing PySpark for building data pipelines in Azure Data Factory to support Machine Learning model deployments.
- Extensive experience with Azure Services such as Data Lake, Data Lake Analytics, SQL Database, Synapse, Databricks, Data Factory, Logic Apps, Function App, and EventHub.
- Working knowledge of Azure cloud components including HDInsight, Blob storage, Data Lake, Storage Explorer, SQL DB, SQL DWH, and CosmosDB.
- Experience in setting up build and deployment automation for Terraform scripts using Jenkins.
- Provisioned highly EC2 instances using Terraform and cloud formation and wrote new plugins to support new functionality in Terraform.
- Implemented robust data governance and security measures, leveraging Azure Active Directory, Key Vault, and Private Link for secure data access and encryption.
- Designed and deployed scalable data architectures on Azure, utilizing services like Azure Databricks, Synapse Analytics, and Azure Data Factory for efficient data processing and analysis.
- Collaborated with cross-functional teams to develop and maintain Azure DevOps pipelines, ensuring seamless integration, testing, and deployment of data-intensive applications.
- Migrated terabytes of data via ETL pipelines using PySpark & Databricks, reducing processing time by 30%.
- 5+ Years of experience in Big Data Analytics using Various Hadoop eco-systems tools and Spark Framework and Azure cloud services using Scala and python as the main programming dialect.
- Architected & managed data workflows with Databricks Jobs & Airflow for timely, monitored ETL across diverse sources.
- Experienced Data Engineer with expertise in utilizing Hadoop ecosystem tools like HDFS, MapReduce, YARN, Spark, Kafka, Hive, Sqoop, Pig, Impala, HBase, Flume, Oozie, and Zookeeper.
- Strong expertise in troubleshooting and performance fine-tuning Spark and Hive applications.
- Proficiency in developing data pipelines using Hive and Sqoop for extracting data from weblogs and storing it in HDFS, along with developing HiveQL for data analytics.
- Extensive experience in automating data ingestion and transformation pipelines using Apache Spark and Airflow.
- Substantial experience in writing MapReduce jobs in Java.
- Extensively dealt with Spark Streaming and Apache Kafka for handling live stream data.
- Experience with real-time data ingestion using Kafka.
- Real-time analytics & decision-making enabled by Apache Kafka integration and PySpark streaming data processing.
- Well-versed in managing large databases including Hive, Oracle, SQL Server, SQL, PL/SQL, and T-SQL.
- Contributed to the development of data warehousing and ETL tools SSIS, Informatica, and PowerBI.
- Develop and manage ELT processes in a Snowflake data warehouse using dbt and Fivetran to support data needs for a SaaS business.
- Communication Skills, Jira, Agile, Cross-team Collaboration
- Experience in Agile and Waterfall methodologies with excellent communication skills for client-facing meetings.

EDUCATION:

TECHNICAL SKILLS:

Azure Services	ADV2, BLOB, ADLS, Azure SQL DB, SQL server, Azure Synapse, Azure Analytics Services, Data bricks, Mapping Dataflow (MDF), Azure data lake (Gen 1/Gen2), Azure Cosmos DB, Azure, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning, App service, Logic apps, Event Grid, Services, Logic apps, Event Grid, Service Bus, Azure Devops, GIT Repository Management, ARM Templates.
Languages	SQL, PL/SQL, Python, HiveQL, Scala, Java.

Web Technologies	HTML, CSS, JavaScript, XML, JSP, Restful, SOAP
Big Data Technologies	HDFS, MapReduce, Hive, Sqoop, Oozie, Zookeeper, Kafka, Apache Spark, Spark Streaming,
Hadoop Distribution	Cloudera, Horton Works
Operating Systems	Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS.
Build Automation tools	Ant, Maven
Version Control	GIT, GitHub.
IDE &Build Tools, Design	Eclipse, Visual Studio.
Databases & Query Language	MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse, Azure Cosmos, Vertica, Teradata. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB, Mongo DB

PROFESSIONAL EXPERIENCE:

Wells Fargo , Irving, TX **Aug 2022 – Present**
Azure Data Engineer

- Supervised ETL processes for sourcing diverse healthcare data streams, including electronic health records (EHRs), claims data, and member enrollment information, as part of optimizing Digital Health Solutions for Blue Cross and Blue Shield Companies.
- Demonstrated proficiency in Azure cloud components including Databricks, Data Lake, Blob Storage, Data Factory, Storage Explorer, SQL DB, SQL DWH, and Cosmos DB.
- Executed Extract Transform and Load operations from Source Systems to Azure Data Storage services employing Azure Data Factory, Databricks, PySpark, Spark SQL, and U-SQL Azure Data Lake Analytics.
- Designed and implemented pipelines in Azure Data Factory to Extract, Transform, and load data from Azure SQL, Blob storage, and Azure SQL Data Warehouse.
- Built streaming ETL pipelines using Spark Streaming to extract data from various sources, transform it in real-time, and load it into a data warehouse such as Azure Synapse Analytics.
- Managed Azure BLOB and Data Lake storage, loading data into Azure SQL Synapse Analytics (DW).
- Developed data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL.
- Analyzed data from Azure data storages using Databricks to derive insights utilizing Spark cluster capabilities.
- Developed a Spark Streaming application to process real-time data from various sources such as Kafka and Azure Event Hubs.
- Successfully ingested data from Azure Blob Storage to Snowflake and then into Palantir Foundry.
- Designed and implemented a real-time data streaming solution utilizing Azure EventHub.
- Conducted performance tuning and optimization activities to ensure optimal performance of Azure Logic Apps and associated data processing pipelines.
- Constructed data infrastructure to support front-end web platforms, data science, and AI applications for a healthcare quality SaaS company using Python, SQL, PySpark, dbt, and Azure Databricks.
- Contributed to the development of designing, developing, and maintaining large data pipelines using Palantir Foundry as an enterprise tool.
- Utilized tools such as Azure Databricks or HDInsight to scale out the Spark Streaming cluster as needed.
- Involved in the complete Big Data flow of the application starting from data ingestion from upstream to HDFS, processing, and analyzing the data in HDFS.
- Experienced in Data Modeling & Data analysis utilizing Dimensional Data Modeling and Relational Data Modeling, Star Schema/Snowflake Modeling, FACT & Dimensions tables, Physical & Logical Data Modeling.
- Provided advice, guidance, and best practices around Snowflake, including guidance on moving data across different environments in Snowflake.
- Acted as a Palantir Foundry Specialist, contributing to the creation of a data pipeline including PySpark and Typescript development.
- Utilized Power BI Gateways for refreshing datasets connected to on-premise platforms ensuring dashboards and reports reflect the latest data.
- Developed comprehensive multiple 'Data Mapping' requirements for sources 'Voyager' and 'Palantir' to target ETL processes.

- Developed Spark API to import data into HDFS from Teradata and created Hive tables.
- Engaged daily with dbt and Snowflake to create both transformational data models and conformed dimensions.
- Created Partitioned and Bucketed Hive tables in Parquet File Formats with Snappy compression.
- Loaded data into Parquet Hive tables from Avro Hive tables.
- Implemented Hive partitioning, bucketing, optimization code through set parameters, and performed different types of joins on Hive tables, implementing Hive SerDe like Avro, JSON

Environment: Azure, Azure Data Factory, Azure Databricks, Azure Blob Storage, Azure SQL Synapse Analytics (formerly Azure SQL Data Warehouse), Azure Data Lake, Azure Event Hub, Azure Logic Apps, Snowflake, Palantir Foundry, HDInsight, HDFS, Hive, Spark SQL, Spark Streaming, Kafka, Python, SQL, PySpark, dbt, Power BI, Teradata, Informatica, Oracle, CI/CD, PL/SQL, UNIX Shell Scripting, Cloudera.

Molina Healthcare,Long Beach, CA **Apr2020 -Aug 2022**
AZURE Data Engineer

- Demonstrated hands-on experience with Azure Cloud Services, Azure Synapse Analytics, SQL Azure, and Azure Data Factory, integrating Snowflake for data processing and analysis.
- Created Batch & Streaming Pipelines in Azure Data Factory (ADF) for Extract, Transform, and Load (ETL) operations, incorporating Snowflake and dbt for data processing and transformation.
- Designed Azure Data Factory (ADF) Batch pipelines for ingesting data from relational sources into Azure Data Lake Storage (ADLS gen2) incrementally, cleansing it, and loading it into Snowflake Delta tables.
- Implemented Azure Logic Apps to automate data loading processes from email attachments to Azure Blob Storage, integrating Snowflake for downstream data processing.
- Developed Spark jobs in Java for indexing data into Azure Functions from external Hive tables in HDFS.
- Built Spark Streaming applications for real-time analytics, leveraging Snowflake and dbt for data processing and transformation.
- Utilized Spark SQL for querying and aggregating real-time data, integrating Snowflake for unified data modeling.
- Developed Spark Streaming applications integrated with event-driven architectures such as Azure Functions or Azure Logic Apps, incorporating Snowflake for data processing.
- Transformed and copied data from JSON files in Data Lake Storage into Azure Synapse Analytics tables using Azure Databricks.
- Conducted performance comparisons between Spark and traditional Big Data technologies like MapReduce and Hive using Scala, leveraging Snowflake for data analysis.
- Applied dbt best practices for sourcing and transforming video streaming subscription data, ensuring a unified data model.
- Created Hive tables, developed custom HiveUDFs, and utilized JSON and XMLSerDe's for loading and analyzing data.
- Migrated ETL processes from Oracle to Hive, leveraging Snowflake for data manipulation and analysis.
- Implemented Kafka for reprocessing failure messages, utilizing offsetid, and analyzed partitioned and bucketed data using HiveQL.
- Developed Sqoop Jobs for loading data from RDBMS to external systems like HDFS and Hive.
- Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation.
- Utilized RDD transformations for filtering data in SparkSQL, transforming dynamic XML data for injection into HDFS.
- Developed CI/CD pipelines using Azure DevOps with GIT and Maven, integrating Snowflake and dbt for continuous deployment and integration.
- Configured Spark Streaming to receive real-time data from Apache Flume and store the stream data.
- Integrated Hive and SQL Contexts with Spark SQL for optimal performance, utilizing GIT for version control in coordination with CI tools, and loading data from UNIX file systems to HDF

Environment: Azure Cloud Platform, Azure Data Factory, Azure Databricks, Azure Blob Storage, Azure Synapse Analytics (formerly SQL Data Warehouse), Azure Logic Apps, Azure DevOps, GIT, Maven, Jenkins, Snowflake Data Warehouse, dbt (Data Build Tool), Spark Streaming, Scala, Kafka, Hive, HDFS, Sqoop, PySpark, Spark-SQL, XMLSerDe, JSON, Apache Flume, Azure Functions, Terraform, Apache Spark, UNIX File System, SQL, CI/CD, Cloudera.

Clairvoyant, Chandler, AZ **Feb2019- Mar2020**
Hadoop developer/Data Engineer

- Contributed to the development of development of data ingestion pipelines using ETL tool, Talend & bash scripting with big data technologies including but not limited to Hadoop, Hive, Spark, Kafka.
- Experience in developing scalable & secure data pipelines for large datasets.
- Gathered requirements for ingestion of new data sources including life cycle, data quality check, transformations, and metadata enrichment.
- Developed data pipeline using Flume, Sqoop, Pig and Java Map Reduce to ingest customer behavioural data into HDFS for analysis.
- Supported data quality management by implementing proper data quality checks in data pipelines.
- Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
- Implemented data streaming capability using Kafka and informatica for multiple data sources.
- Involved in SQOOP implementation which helps in loading data from various RDBMS sources to Hadoop systems
- Worked with multiple storage formats (Avro, Parquet) and databases (Hive) Azure SQL.
- Optimizing query performance in Hive using bucketing and partitioning techniques
- Creating and managing partitions and buckets in Hive tables.
- Responsible for maintaining and handling data inbound and outbound requests through big data platform.
- Used Sqoop to transfer data between relational databases and Hadoop.
- Knowledge on implementing the JILs to automate the jobs in production cluster.
- Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
- Contributed to the development of analysing and resolving the production job failures in several scenarios.
- Implemented UNIX scripts to define the use case workflow and to process the data files and automate the jobs.

Environment: Spark, AZURE SQL, Python, HDFS, Hive, Sqoop, Scala, Kafka, Shell scripting, Linux, Eclipse, Git, Oozie, Informatica, Agile Methodology.

Education: Masters in computer science form the University of Dayton.