A blue and white sign

Description automatically generated**Shantan Jonnalagadda**

[shantanj091@gmail.com](mailto:shantanj091@gmail.com) **|**+1(227)254-9970

**Professional Summary**

* Around 13 years of experience as a Data Engineer specializing in cloud-based and big data solutions.
* Experience with all stages of the SDLC and Agile Development model right from the requirement gathering to Deployment and production support.
* Good exposure to Python programming.
* In-depth knowledge of Snowflake Database and Table structures.
* Hands - on experience in Azure Cloud Services: Azure Synapse Analytic, SQL Azure, Data Factory, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake.
* Excellent experience on AWS cloud services: (Amazon Redshift and Data Pipeline).
* Skill experience in installing, configuring and using Apache Hadoop ecosystems such as Map Reduce, Hive, Pig, Sqoop, Spark, Kafka and Oozie.
* Good knowledge of Python and corresponding libraries with experience of development for several platforms (Windows, UNIX). Experience in Dimensional Data Modeling experience using Data modeling, Relational Data modeling, ER/Studio, Erwin, and Sybase Power Designer.
* Expert in writing SQL queries and optimizing the queries in Oracle, SQL Server and Teradata.
* Working experience in PL/SQL, SQL\*Plus, Stored Procedures, Functions & Packages.
* Excellent technical and analytical skills with clear understanding of design goals and development for OLTP and dimension modeling for OLAP.
* Depth knowledge in Data Warehousing concepts with emphasis on ETL and ELT
* Excellent Programming skills at a higher level of abstraction using Scala and Python.
* Good working experience on Spark (spark streaming, spark SQL) and Scala.
* Experience in Data transformation, Data Mapping from source to target database schemes, Data Cleansing procedures.
* Experience in **Cloud computing on Google Cloud Platform** with various technology like Data-flow, Pub/Sub, Big Query and all related tools.
* Strong experience with big data processing using Hadoop technologies Map Reduce, Apache Spark, Hive and Pig.
* Data importing and exporting by using Sqoop from HDFS to Relational Database Systems and vice-versa.
* Developed multi-cloud strategies in better using GCP (for its PAAS) and Azure (for its SAAS).
* Experience in transferring the data using Informatica tool from AWS-to-AWS Redshift
* Excellent knowledge and extensively using NOSQL databases (HBase).
* Proficient knowledge in Designing and implementing data structures and commonly used data business intelligence tools for data analysis.
* Good experience in creating and designing data ingest pipelines using technologies such as Kafka.
* Experience in importing/exporting data using Sqoop into HDFS from Relational Database Systems and vice - versa.
* Good knowledge in Data Analysis, Data Validation, Data Cleansing, Data Verification and identifying data mismatch.
* Strong experience in Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export.
* Expertise in SQL Server Analysis Services (SSAS) and SQL Server Reporting Services (SSRS) tools.
* Must be able to bring best of GCP technologies and leverage open source like terraform, and chef recipes to build the solution
* Experience designing security at both the schema level and the accessibility level in conjunction with the DBAs.

**Education& Certification:**

* Bachelor of Technology in EIE from Jawaharlal Nehru Technological University
* Certified Azure Data Engineer

**Experience:**

**US Bank, Minneapolis, MN 02/25 to Present**

**Sr. Data Engineer**

**Description:** U.S. Bank is a leading financial institution focused on secure and efficient banking services. I built a payment processing system to handle high volume transactions smoothly, ensuring fast and reliable transfers for customers. The project involved integrating data from multiple payment channels to reduce delays and errors. It now supports seamless operations for millions of daily transactions.

**Responsibilities:**

* Designed efficient ETL pipelines using Apache Spark and Apache Airflow to process large-scale datasets.
* Developed data lakes on Azure Data Lake Storage, ensuring secure and optimized data storage and retrieval.
* Utilized Azure Databricks for advanced data processing, machine learning model integration, and real-time analytics.
* Orchestrated data workflows with Azure Data Factory to automate data movement and transformation across multiple sources.
* Built and optimized data models in Azure Synapse Analytics to support business intelligence and reporting needs.
* Integrated Hadoop ecosystem tools like Hua and HBase for querying and managing large datasets in distributed environments.
* Implemented real-time data streaming solutions using Apache Kafka and Azure Event Hubs for low-latency data ingestion.
* Ensured data quality and governance by utilizing Azure Purview for metadata management and compliance monitoring.
* Collaborated with cross-functional teams to deploy machine learning models on Azure Machine Learning, enhancing predictive analytics.
* Monitored and optimized data pipeline performance using Azure Monitor and Log Analytics to ensure high availability and reliability.
* Configured and managed Azure Kubernetes Service (AKS) to deploy containerized data processing applications, improving operational flexibility.
* Developed custom data ingestion scripts using Python and PySpark to handle diverse data formats and sources.
* Utilized Apache NiFi for automating data flows between heterogeneous systems, enhancing data integration efficiency.
* Implemented data security measures using Azure Key Vault to manage encryption keys and sensitive credentials.
* Created interactive dashboards with Power BI, integrating Azure data sources to provide actionable insights for stakeholders.

**Environment:**, Logic Apps, Functional App, Snowflake, SparkSQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Jenkins, Kafka, Power Bi, Spark Streaming, Azure Databricks, Data Factory, Snowflake Data Warehouse

**Infosys/Truist |India/USA 02/2023 to 12/2024**

**Sr. Data Engineer**

**Description:** Truist is a leading financial institution focused on enhancing customer experiences through advanced analytics. The Customer 360° Data Platform for Personalized Banking Services integrates data from multiple sources to create a unified customer profile, enabling tailored financial products and services. This platform uses big data technologies, AI driven insights, and batch processing to enhance customer engagement.

**Responsibilities:**

* 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 the customer usage patterns.
* Designed and implemented end-to-end data pipelines using Azure Data Factory to facilitate efficient data ingestion, transformation, and loading (ETL) from diverse data sources into Snowflake data warehouse.
* Extract Transform and Load data from sources Systems to Azure Data Storage services using a combination of Azure Data factory, T-SQL. Spark SQL and U-SQL Azure Data Lake Analytics. Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage. Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Identified, evaluated, and documented potential data sources in support of project requirements within the assigned departments as per agile methodology.
* Developed efficient data models and pipelines using Fabric, optimizing for analytical queries and business intelligence reporting.
* Developed API proxies in Apigee to route and manage traffic between external clients and backend services, ensuring robust API security and performance.
* Developed and deployed the code to different environments using Azure DevOps CI/CD Pipelines.
* Involved with impact assessment in terms of schedule changes, dependency impact, code changes for various change requests on the existing Data Warehouse applications that running in a production environment. Developed Data Mapping, Data Governance, Transformation, and cleansing rules.
* Utilized Flink’s APIs for complex data transformations, aggregations, and joins in real-time streaming and batch applications.
* Integrated Drools rule engine with enterprise applications to execute real-time rule-based processing and validation.
* Designed end-to-end ETL workflows using Informatica PowerCenter to extract, transform, and load data from multiple sources into data warehouses and data marts.
* Developed ETL pipelines using Bash scripts to extract, transform, and load large datasets efficiently.
* Integrated GraphQL with various data sources and databases to fetch and manipulate data efficiently.
* Implemented best practices for data storage, considering both structured and unstructured data.
* Developed robust data pipelines to ingest and integrate data from diverse sources into Azure Data Lake Storage or Azure SQL Data Warehouse.
* Utilized PMMT for monitoring data pipeline performance and identifying bottlenecks in ETL processes to optimize data workflows.
* Built RESTful APIs using Node.js to facilitate seamless data access between front-end applications and backend data systems.
* Built stream processing applications using Apache Flink or Spark Streaming to perform complex transformations, aggregations, and computations on continuous data streams.
* Designing GraphQL schemas tailored to specific application requirements, ensuring efficient data fetching and manipulation.
* Utilized Power Query for data cleaning, transformation, and preparation before visualizing in Power BI.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Implemented real-time and batch processing techniques for efficient data movement.
* Utilized Azure Databricks & Azure HD Insight for large-scale data processing and transformation.
* Integrated Apache NiFi with Kafka for real-time streaming and event-driven architectures.
* Implemented ETL (Extract, Transform, Load) processes to ensure data quality and consistency.
* Analyze, design and build Modern data solutions using Azure Pass service to support visualization of data.
* Implement one time Data Migration of Multistate level data from SQL server to Snowflake by using Python and SnowSQL.
* Developed reusable, modular Terraform templates to standardize infrastructure deployments and enhance maintainability.
* Designed and implemented metadata repositories to store and manage technical, business, and operational metadata across data pipelines
* Working in ETL methodology for supporting Data Analysis, Extraction, Transformations and Loading, in a corporate-wide-ETL Solution using Ab Initio.
* Expertise in using Azure DevOps for code check-in, creating pull request and configuring build and release definitions.
* Extensively worked on Python and build the custom ingest framework and worked on Rest API using python.
* Developed data warehouse model in snowflake for over 100 datasets using where cape.
* Expertise in data loads, data transformations, and fixing incorrect data in the Oracle and PostgreSQL tables.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline to Extract, Transform and load data from different sources like Azure SQL, Bold storage, Azure SQL Data warehouse, write-back tool and backwards.
* 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 the customer usage patterns.
* Performed data blending and transformation using Tableau Prep to ensure consistent and clean data for analysis.
* Analyzed existing systems and propose improvements in processes and systems for usage of modern scheduling tools like Airflow and migrating the legacy systems into an Enterprise data lake built on Azure Cloud.
* Developed data pipeline using Kafka, Flume, Pig and Map reduce to ingest customer behavioral data and financial histories into HDFS for analysis.
* Used JavaScript libraries (e.g., D3.js, Chart.js) for building dynamic, interactive data visualizations and dashboards to present data insights.
* Designed and implemented scalable data pipelines using Apache NiFi, handling high-volume data ingestion from multiple sources.
* Experienced in writing Spark Applications in Scala and Python.
* Analyzed large and critical datasets using HDFS MapReduce, Kafka, Spark, HBase, Hive, Hive UDF and Spark.
* Used Kafka consumer’s API in Scala for consuming data from Kafka topics.
* Analyzed the SQL scripts and designed the solution to implement using Pyspark.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines
* Developed Python code to gather the data from HBase (Cornerstone) and designs the solution to implement using PySpark.
* Responsible for estimating cluster size, monitoring, and troubleshooting the Spark Databricks cluster.
* Involved in ETL, Data Integration and Migration by writing Pig scripts.
* Exported the analyzed data to Teradata using Sqoop for visualization and to generate reports for the BI team.
* Migrated the computational code in HQL to PySpark.
* Optimized NiFi workflows to improve data processing efficiency and reduce latency
* Work related to downloading BigQuery data into pandas or Spark data frames for advanced ETL capabilities
* Completed data extraction, aggregation, and analysis in HDFS by using PySpark and store the data needed to Hive.
* Exposure on usage of Apache Kafka develop data pipeline of logs as a stream of messages using producers and consumers.
* Expertise in snowflake to create and Maintain Tables and views.
* Sound knowledge in programming Spark using Scala.
* Populated HDFS and HBase with huge amounts of data using Apache Kafka.
* Experienced in working with various kinds of data sources such as Teradata and Oracle. Successfully loaded files to HDFS from Teradata, and load loaded from HDFS to Hive and Impala.
* Experience on cloud versioning technologies like Github.
* Created Build and Release for multiple projects in production environment using Visual Studio Team Services.

**Environment:** PySpark, Python, Azure Databricks, Azure Data Lake Storage (ADLS Gen2), Azure Data Factory (ADF), Azure Synapse Analytics (ASA), SQL Server Management Studio (SSMS), Azure Data Studio, Azure Data Lake Analytics, POSTMAN, BLOB Storage, AWS S3, Azure DevOps, Azure Repos, Azure ML.

**Wipro |Corning Incorporated 01/2020 to 02/2023**

**Sr. Data Engineer**

**Description:** Corning is a global leader in materials science, specializing in glass, ceramics, and advanced optical solutions. As a Data Engineer, I contributed to the Smart Manufacturing Data Platform, developing data pipelines to process IoT sensor data from factory equipment.

**Responsibilities:**

* Highly Involved into Data Architecture and Application Design using Cloud and Big Data solutions on Microsoft Azure.
* Leading the effort for migration of Legacy-system to Microsoft Azure cloud-based solution. Re-designing the Legacy Application solutions with minimal changes to run on cloud platform.
* 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 and controlling and granting database access.
* Designed and implement data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from these sources into a centralized data lake or data warehouse.
* Configured Fabric components to support real-time data streaming and event-driven architectures for processing large-scale data.
* Built the data pipeline using Azure Service like Data Factory to load the data from Legacy SQL server to Azure Data Base using Data Factories, API Gateway Services, SSIS Packages, Talend Jobs and Python codes.
* Creating build and release pipelines using Azure DevOps.
* Integrated stream processing with batch processing to create hybrid pipelines for comprehensive data analytics and machine learning model training.
* Implemented MongoDB Change Streams for real-time data processing and event-driven architectures.
* Built Bash scripts to monitor and log data pipeline performance, ensuring timely detection and resolution of failures.
* Built Azure Web Job for Product Management teams to connect to different APIs and sources to extract the data and load into Azure Data Warehouse using Azure Web Job and Functions.
* Integrated ODM with data systems to enable dynamic rule evaluation and automated decision-making within ETL workflows.
* Developed Node.js scripts for ingesting data from external APIs and databases into internal data systems for further analysis.
* Used Snowflake to store the transformed data which is consumed by data scientists, reporting layer etc.
* Expertise in handling errors and exceptions within GraphQL APIs, ensuring robustness and reliability.
* Optimized GraphQL queries for performance, including techniques such as batching, caching, and query optimization.
* Involved in Agile Development process (Scrum and Sprint planning).
* Developed real time data processing applications by using Scala and Python and implemented Apache Spark Streaming from various streaming sources like Kafka and JMS.
* Configured OAuth, API key validation, and JWT token authentication in Apigee to secure APIs, protecting sensitive data and preventing unauthorized access.
* Configured NiFi Provenance Data for monitoring, debugging, and troubleshooting data flow issues.
* Developed SQL-based transformation workflows in dbt to clean, aggregate, and enrich raw data for analytics.
* Integrated Flink with Kafka, Cassandra, and HDFS for seamless data ingestion and processing.
* Used Python and Django creating graphics, XML processing, data exchange and business logic implementation.
* Implement real-time data processing solutions using Azure Stream Analytics to enable near-instantaneous insights from streaming data sources.
* Configured and managed Terraform state files securely using remote backends like S3, GCS, or Azure Blob Storage.
* Developed and managed Drools knowledge bases (KIE sessions) to ensure efficient rule execution and scalability.
* Led migration efforts from legacy ETL tools to Informatica PowerCenter, ensuring minimal downtime and data accuracy.
* Ensure the scalability and reliability of real-time data pipelines.
* Have good experience working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Design and development software products targeted towards financial institutions using Ab Initio.
* Heavily involved in testing Snowflake to understand best possible way to use the cloud resources.
* Performed end-to-end delivery of PySpark ETL pipelines on Azure-data-bricks to perform the transformation of data orchestrated via Azure Data Factory (ADF) scheduled through Azure automation accounts and trigger them using Tidal Scheduler.
* Spearheaded the migration initiative from Control-M to Apache Airflow, ensuring a seamless transition of workflow orchestration for critical business processes.
* Created advanced calculated fields and parameters in Tableau to enable dynamic and custom visualizations.
* Developed and maintained comprehensive documentation outlining the migration plan, including task schedules, dependencies, and fallback strategies, resulting in minimal downtime during the migration process.
* Developed server-based web traffic using RESTful API's statistical analysis tool using Flask, Pandas.
* Expertise in creating and deploying tables in Oracle & PostgreSQL.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Design & implement migration strategies for traditional systems on Azure.
* Knowledge on PySpark and used Hive to analyze sensor data and cluster users based on their behavior in the events.
* Implemented Proof of concepts for SOAP & REST APIs.
* Configured Flink applications for high availability and fault tolerance using checkpointing and state recovery features.
* Creating Azure Databricks notebooks using SQL, Python and automated notebooks using jobs.
* Developed Spark Python modules for machine learning & predictive analytics in Hadoop.
* Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL.
* Used various sources to pull data into Power BI such as SQL Server, Excel, Oracle, SQL, Azure etc.
* Worked on migrating Map Reduce programs into Spark transformations using Spark, and Scala, initially done using python (PySpark).
* Designed, developed, and maintained ETL (Extract, Transform, Load) pipelines using Azure Data Factory (ADF).
* Developed PySpark programs and created the data frames and worked on transformations.
* Involved in loading data from Linux file systems, servers, web services using Kafka producers and partitions.
* Applied Kafka custom encoders for custom input format to load data into Kafka Partitions.
* Implement POC with Hadoop. Extract data with Spark into HDFS.
* Used Spark SQL with Scala for creating data frames and performed transformations on data frames.
* Creating Spark clusters and configuring high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Optimized Drools rule execution by structuring rule groups, utilizing salience, and leveraging rule indexing for better efficiency.
* Implemented automated testing frameworks using Python and Azure DevOps, ensuring data quality and validating data pipelines through unit tests, integration tests, and end-to-end testing.
* Worked on Snowflake Schema, Data Modeling and Elements, and Source to Target Mappings, Interface Matrix and Design elements. Performed data quality issue analysis using Snow SQL by building analytical warehouses on Snowflake.
* Developed Spark applications using Scala for easy Hadoop transitions.
* Optimized the code using PySpark for better performance.
* Integrated dbt projects with Git for version control, ensuring collaborative development and change tracking across data pipelines.
* Integrated Terraform with tools like Sentinel or OPA (Open Policy Agent) to enforce compliance policies and security standards.
* Worked on Spark streaming using Apache Kafka for real time data processing.
* Experienced in optimizing Hive queries, joins to handle different data sets.
* Implemented fault-tolerant NiFi clusters, ensuring high availability and disaster recovery.
* Involved in ETL, Data Integration and Migration by writing Pig scripts.
* Deployed and managed containerized applications on Kubernetes clusters using Helm charts, YAML manifests, or Kustomize.
* Developed MapReduce programs to cleanse the data in HDFS obtained from heterogeneous data sources.
* Designed and implemented MongoDB and associated RESTful web service.
* Utilized NiFi Expression Language for dynamic routing, filtering, and custom transformations.
* Involved in writing test cases and implement test classes using MR Unit and mocking frameworks.
* Developed Shell, Perl and Python scripts to automate and provide Control flow to Pig scripts.
* Used Talend tool to create workflows for processing data from multiple source systems.
* Worked on Jenkins continuous integration for deployment of project and deployed the project into Jenkins using GIT version control system.

**Environment:** Azure Databricks, Data Factory, Snowflake Data Warehouse, Logic Apps, Functional App, Snowflake, SparkSQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Jenkins, Kafka, Power Bi, Spark Streaming.

**Wipro| Indian Bureau of Mines 01/2017 to 01/2020**

**Sr. Data Engineer**

**Description:** Indian Bureau of Mines is a regulatory agency overseeing mining activity and mineral resource management in India. The Geospatial Data Processing for Mine Mapping and Compliance project focused on integrating GIS and satellite imagery to map mining sites, monitor land use, and ensure regulatory compliance.

**Responsibilities:**

* Developed detailed project plans to manage and streamline data conversion and migration, successfully moving data to Snowflake databases while minimizing downtime and ensuring data integrity.
* Designed and implemented large-scale Business Intelligence (BI) solutions using Azure Data Platform services, including Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL Data Warehouse, HDInsight, Databricks, and NoSQL databases, which improved insights and performance in complex data environments.
* Implemented and optimized custom Azure Data Factory pipelines, automating data ingestion and transformation workflows for real-time processing and delivery across the Azure ecosystem.
* Conducted ad-hoc analysis and data transformations using Azure Data Lake Analytics and HDInsight, enabling timely insights and data-driven decisions for business stakeholders.
* Migrated and integrated data from Azure Data Lake to Azure Data Warehouse using PolyBase, setting up external tables with optimized configurations and executing scheduled data loads for high-volume ingestion.
* Managed data ingestion from various sources into Azure SQL Data Warehouse and transformed data within Azure SQL Database, optimizing data storage solutions in Azure Data Lake.
* Implemented real-time data streaming using Spark Streaming and Kafka to process and store streaming data in HDFS, ensuring high availability and scalability for real-time applications.
* Designed MongoDB-based data pipelines, optimizing data storage and retrieval.
* Developed and maintained data pipelines on Azure Analytics platforms using Azure Databricks, focusing on automating workflows, ensuring data quality, and optimizing performance for large-scale processing.
* Designed and implemented purging routines on Azure SQL Server and Azure Blob Storage, creating scripts to maintain storage efficiency and manage large datasets
* Used Bash and SQL to automate database maintenance tasks such as indexing, backup, and archiving.
* Created Azure Databricks notebooks to perform complex data transformations, cleansing, and apply business logic for analytics, streamlining operations and improving accuracy.
* Engineered complex data pipelines using Azure Databricks and Azure Data Factory to integrate data from various sources and create a unified data lake for seamless flow across the organization.
* Developed high-performance Hive SQL scripts for creating complex tables, utilizing partitioning, clustering, and bucketing strategies to optimize query execution in big data environments.
* Ingested and processed data through Hadoop HBase into Spark RDD for in-memory computations, continually optimizing User Defined Functions (UDFs) and data structures for enhanced query performance.
* Authored PySpark scripts for secure data encryption, using hashing algorithms to ensure privacy and security of sensitive client data.
* Leveraged Apache Airflow for creating, scheduling, and monitoring complex ETL pipelines, establishing automated workflows to improve the efficiency and reliability of data processing.
* Worked on large-scale data migration projects utilizing SQL, Azure Storage, Azure Data Factory, SSIS, and PowerShell, ensuring robust and efficient data transfers.
* Transformed Hive/SQL queries into Spark RDDs using Python and Scala, enhancing processing speeds and optimizing for large-scale distributed environments.

**Environment**: Python, Hadoop, Spark, Spark SQL,Streaming, PySpark, Hive, Scala, MapReduce, HDFS, Kafka, Sqoop, HBase, MS Azure, Blob Storage, Data Factory, Data Bricks, SQL Data Warehouse, Apache Airflow,Snowflake, Oracle, MySQL,UNIX Shell Script, Perl, PowerShell, SSIS, Power BI.

**Wipro|** **Symetra Financial 04/2015 to 01/2017**

**Sr. Big Data Engineer**

**Description:** Symetra Financial is a leading provider of insurance, benefits, and annuities. The Investment Portfolio Analysis project focused on building a robust data pipeline to aggregate, process, and analyze financial data from multiple sources, ensuring real time risk assessment and portfolio optimization.

**Responsibilities:**

* Participated in gathering requirements, analyzing systems, and providing development and testing effort estimates.
* Analyzed enterprise data integration and developed functional specifications for building reporting systems.
* Migrated on-premises applications to AWS, utilizing services like EC2 and S3 for data processing and storage, and managed Hadoop clusters using AWS EMR.
* Performed data transformations using Apache Spark and stored results in HDFS and Snowflake.
* Designed and developed ETL processes using AWS Glue to migrate data from S3 and file formats like JSON, Parquet, Text Files into AWS Redshift.
* Managed real-time data analytics using Spark Streaming, Kafka, and Flume, configuring Spark to handle streaming from Kafka and storing data in HDFS.
* Transformed and cleansed input data extracted from external sources using Apache Spark.
* Developed Bash scripts to orchestrate Spark, Hadoop, and Kafka jobs, enhancing big data processing efficiency.
* Created DataStage jobs using stages such as Transformer, Aggregator, Join, Lookup, Data Set, and Funnel for data transformation.
* Designed and monitored ETL batch processes using Apache Airflow, loading data into Snowflake for analysis.
* Built ETL pipelines for data ingestion, transformation, and validation on AWS, ensuring compliance with data stewardship.
* Scheduled and validated ETL jobs using Airflow scripts in Python, and managed task dependencies using Lambda.
* Extracted, filtered, and transformed data in pipelines using PySpark for optimized performance.
* Monitored servers using AWS CloudWatch, ensuring data retrieval from workflows.
* Utilized Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats to analyze customer usage patterns.
* Developed scripts to migrate relational database (MySQL, PostgreSQL) data to MongoDB, ensuring schema compatibility.
* Designed sharding and indexing strategies to enhance MongoDB query performance for large-scale datasets.
* Optimized queries using Hive and Spark with advanced techniques like window functions.
* Estimated cluster size and handled monitoring and troubleshooting of Spark Databricks clusters, while automating processes using UNIX shell scripting.
* Designed dashboards and created visualizations using Tableau for business insights.

**Environment:** AWS (EC2, S3, Redshift, Glue, EMR, CloudWatch, Lambda), Apache Spark, Kafka, Flume, HDFS, Snowflake, PySpark, Apache Airflow, Databricks, Hive, DataStage, UNIX, Docker, Jenkins, Tableau.

**Zensar Technologies|** **Bangalore, India 04/2012 to 04/2015**

**Software Developer**

**Description:** Zensar is a global IT services company specializing in digital transformation and data solutions. As part of the Big Data Processing and Analytics project, I developed and optimized Hadoop-based ETL pipelines to process large-scale structured and unstructured data efficiently.

**Responsibilities:**

* + Involved in creating data ingestion pipelines for collecting clinical trial and subject’s data from various external sources like FTP Servers and S3 buckets.
  + Involved in migrating existing Teradata Datawarehouse to AWS S3 based data lakes.
  + Involved in migrating existing traditional ETL jobs to Spark and Hive Jobs on new cloud data lake.
  + Wrote spark applications for performing various de-normalization of the datasets and creating a unified data analytics layer for downstream teams.
  + Developed series of data ingestion jobs for collecting data from multiple channels and external applications in Java.
  + Primarily responsible for fine-tuning long running spark applications, writing custom spark UDFs, troubleshooting failures etc.,
  + Involved in building a real time pipeline using Kafka and Spark streaming for delivering event messages to downstream application team from an external rest-based application.
  + Developed UDF's using both Data frames/SQL and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop.
  + Worked on utilizing AWS cloud services like S3, EMR, Redshift, Athena, and Glue megastore.
  + Used broadcast variables in spark, effective & efficient Joins, caching, and other capabilities for data processing.
  + Developed various components on java using spring batch that can be used in ETL batch processing jobs.

**Environment:**  AWS EMR, Java, Lambda, IAM, Spark, Hive, Sqoop, Kafka