Sohit Lama

Email: b4a-z3t-f0f@mail.dice.com
Phone: 9292260227

Summary

Previous	
Preferred	Data Engineer
Location	Queens, NY, US
Desired Work Settings	Remote or On-Site or Hybrid
Willing to Relocate	No
Work Authorization(s):	Authorized to work in the United States on a full-time basis.
Employment Type	Full-time Part-time Contract - Corp-to-Corp Contract - Independent Contract - W2 Contract to Hire - Corp-to-Corp Contract to Hire - Independent Contract to Hire - W2
Total Experience	Unspecified
Education	Unspecified
Profile Source	Dice
Profile Downloaded	Wednesday, February 26, 2025

Sohit Lama
Phone:
Email:
LinkedIn: https://www.linkedin.com/in/sohit-lama-1bba5959

Summary:

- Expert Data Engineer with 13+ years of experience in Azure development, Azure Data Factory, Azure Data Lake, T - SQL, and Data bricks.
- Strong experience using Spark RDD API, Spark Data frame/Dataset API, Advanced SQL, Spark - SQL, ANSI SQL, SQL Database Tuning and Spark ML frameworks for building end to end data pipelines.
- Hands on expertise with AWS Databases such as RDS(Aurora), Redshift, DynamoDB and Elastic Cache (Memcached & Redis).
- Experience in creating and executing Data pipelines in GCP and AWS platforms.
- Strong experience building Spark applications using Pyspark and python as programming language.
- Good experience troubleshooting and fine-tuning long running spark applications.
- Hands on experience in GCP, Big Query, GCS, cloud functions, Cloud dataflow, Pub/Sub, cloud shell, GSUTIL command- line utilities, Data Proc.
- Extensive hands-on experience tuning spark Jobs.
- Experienced in working with structured data using Hive QL and optimizing Hive queries.
- Wrote AWS Lambda functions in python for AWS's Lambda which invokes python scripts to perform
- various transformations and analytics on large data sets in EMR clusters.
- Experience in working with real time streaming pipelines using Kafka and Spark-Streaming.
- Strong experience working with Hive for performing various data analysis.
- Developing and migrating on-premises databases to Azure Data Lake stores using Azure Data Factory.
- Proficient knowledge and hand on experience in writing shell scripts in Linux.
- Hands on experience in GCP, Big Query, GCS, cloud functions, Cloud dataflow, Pub/Sub, cloud shell, GSUTIL command- line utilities, Data Proc.
- Adequate knowledge and working experience in Agile and Waterfall Methodologies.
- Excellent understanding and knowledge on NOSQL databases like HBase, Cassandra and Mongo DB.
- Proficient in developing Sqoop scripts for the extractions of data from various RDBMS databases into HDFS.
- Good working experience on different file formats like PARQUET, TEXTFILE, AVRO, ORC and different compression codecs GZIP, SNAPPY.
- Excellent Knowledge in understanding Big Data infrastructure, distributed file systems - HDFS, parallel processing – Map Reduce framework and complete Hadoop ecosystem - Hive, Pig, Sqoop, Spark, Kafka, Hbase, NoSQL, Oozie and Flume.
- Configured and administered Azure like Resource Group, Storage Account, Blob Storage, Delta Lake, Cluster config, Event Hub, Cosmos DB for different zones in development, testing and production environments.
- Experience in working with Azure Code Pipeline and creating Cloud Formation JSON templates to create custom sized VPC & migrate a production infrastructure into an Azure.
- Expertise in using major components of Hadoop ecosystem components like HDFS, YARN, Map Reduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Spark Streaming, Flume, Oozie, Zookeeper, Hue.
- Good knowledge in Database Creation and maintenance of physical data models with Oracle, Teradata, Netezza, DB2, Mongo DB, HBase and SQL Server databases.
- Deep understanding of Map Reduce with Hadoop and Spark. Good knowledge of Big Data ecosystem like Hadoop 2.0 (HDFS, Hive, Pig, Impala), Spark (Spark SQL, Spark MLlib, Spark Streaming).
- Experienced in writing complex SQL Queries like Stored Procedures, triggers, joints, and Sub quires.
- Interpret problems and provide solutions to business problems using data analysis, data mining, optimization tools, and machine learning techniques and statistics.
- Large scale Hadoop environments build and support including design, configuration, installation, performance tuning and monitoring.
- Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales, PivotTables and OLAP reporting.
- Experience in creating and executing Data pipelines in GCP and AWS platforms.
- Experienced with JSON based RESTful web services, and XML/QML based SOAP web services and worked on various applications using python integrated IDEs like Sublime Text and PyCharm

- Developed web-based applications using Python, DJANGO, QT, C++, XML, CSS3, HTML5, DHTML, JavaScript and jQuery.
- Proficient with Apache Spark ecosystem such as Spark, Spark Streaming using Scala and Python.
- Having hands on experience in versioning using bit bucket.
- Hands on Experience in Spark architecture and its integrations like Spark SQL, Data Frames and Datasets APIs.
- Extensive experience in migrating on premise Hadoop platforms to cloud solutions using AWS and Azure.
- Experienced in transporting, and processing real time event streaming using Kafka and Spark Streaming.
- Hands on experience with importing and exporting data from Relational databases to HDFS, Hive and HBase using Sqoop.
- Working with Bootstrap twitter framework to Design single page application.
- Strong working knowledge developing Cross Browser Compatibility (IE, Firefox, Safari, Chrome etc.) for dynamic web applications.
- Experienced in automating end to end data pipelines using Oozie workflow orchestrator.
- Experienced in working with Cloud era, Horton works and AWS big data services.
- Strong experience using and integrating various AWS cloud services like S3, EMR, Glue Meta store, Athena, and Redshift into the data pipelines.
- Strong experience in performance tuning & index maintenance.
- Detailed exposure on Azure tools such as Azure Data Lake, Azure Data Bricks, Azure Data Factory, HDInsight, Azure SQL Server, and Azure DevOps.
- Experienced in requirement analysis, application development, application migration and maintenance using Software Development Lifecycle (SDLC) and Python/Java technologies.
- Good Understanding of Azure Big data technologies like Azure Data Lake Analytics, Azure Data Lake Store, Azure Data
- Factory and created POC in moving the data from flat files and SQL Server using U-SQL jobs.
- Deploying VM's, Storage, Network and Resource Group through Azure Portal.
- Creating Storage Pool and Stripping of Disk for Azure Virtual Machines. Backup, Configure and Restore Azure Virtual Machine using Azure Backup.
- Expertise on Talend Data Integration suite and big data Integration Suite for Design and development of ETL/Big data code and Mappings for Enterprise DWH ETL Talend Project.

Skills:

- **Big Data Technologies:** HDFS, Hive, Map Reduce, Pig, Hadoop distribution, and HBase, Spark, Spark Streaming, Kafka.
- **Cloud Services:** AWS (EC2, S3, EMR, RDS, Lambda, Cloud Watch, Auto scaling, Redshift, Cloud Formation, Glue etc.), Azure(Data bricks, Azure Data Lake, Azure HDInsight)
- **Databases:** Oracle, MySQL, SQL Server, Mongo DB, Dynamo DB, Cassandra, Snowflake.
- **Programming Languages:** Python, Pyspark, Shell script, Perl script, SQL, Java.
- **Tools:** PyCharm, Eclipse, Visual Studio, SQL*Plus, SQL Developer, SQL Navigator, SQL Server Management Studio, Eclipse, Postman.
- **Cloud Tech:** Azure and AWS and GCP
- **Version Control:** SVN, Git, Git Hub, Maven.
- **Operating Systems:** Windows 10/7/XP/2000/NT/98/95, UNIX, LINUX, OS
- **Visualization/ Reporting:** Tableau, ggplot2, matplotlib
- **Big Data Ecosystemts:** HDFS, Map-Reduce, Hive, Pig, Sqoop, Flume, Zookeeper, Spark, Kafka, Spark, Hbase.
- **Deployment Tools:** Git, Jenkins, Terra form and Cloud Formation
- **Cloud Technologies:** Azure Analysis Services, Azure SQL Server, Dynamo DB, Step Functions, Glue, Athena, Cloud Watch, Azure Data Factory, Azure Data Lake, Functions, Azure SQL Data Warehouse, Data bricks and HDInsight
- **Data Visualization:** Tableau, BI Reports, Dremio.

Education: Bachelors in Computer Science, Vidhya Deep Academy, Nepal 2011.

Experience:

Cardinal Health, New York, NY **Sep 21 – Present**
Lead Data Engineer

Responsibilities:

- Designed and implemented scalable, secure, and high-performance data pipelines for healthcare analytics platforms.
- Worked closely with the business intelligence team to design and develop Power BI dashboards and reports.
- Spearhead the design and development of scalable data pipelines using Scala, Apache Spark, and Kafka to support data processing workflows for real-time analytics and batch jobs.
- Led the design and implementation of scalable data architecture using NoSQL databases (MongoDB, Cassandra) to support high-throughput, low-latency data processing applications.
- Architected and implemented scalable data pipelines leveraging Google Kubernetes Engine (GKE) for container orchestration, ensuring high availability and fault tolerance.
- Implemented Data Lakes on Azure Blob Storage, storing both structured and unstructured data for advanced analytics.
- Provided mentoring and training to junior data engineers on Informatica, data modeling, and performance optimization techniques.
- Spearheaded the migration of on-premises ETL workflows to Talend Cloud, ensuring zero data loss and enhanced scalability.
- Led the migration of legacy databases to MongoDB, ensuring minimal downtime and seamless data integration with existing systems.
- Lead and manage a team of data engineers in the design, deployment, and optimization of containerized data solutions using Docker and Kubernetes, ensuring scalability and efficient resource utilization across development, staging, and production environments.
- Spearheaded the design and deployment of a cloud-native data platform on OpenShift, providing a scalable and high-performance architecture for data pipelines and analytics.
- Architected Terraform-based infrastructure for automated provisioning of cloud resources (AWS, GCP) and deployment of data services, reducing manual effort by 40%.
- Designed and deployed an AI/ML solution to predict equipment failures using sensor data.
- Developed and maintained DBT models, documentation, and tests to ensure consistent and accurate data transformations.
- Designed and managed Kubernetes clusters on AWS, using Helm for deployments and Kubernetes networking strategies to ensure high availability and fault tolerance.
- Used various AWS services including S3,EC2, AWS Glue, Athena, RedShift,
- Created real-time data streaming pipelines using GCP Pub/Sub and Dataflow for financial data.
- Experience in GCP Dataproc, GCS, Cloud functions, BigQuery and moving data between GCP and Azure using Azure Data Factory.
- Developed and optimized ETL workflows with Scala and Apache Spark to process data at scale, improving data processing time by 30%.
- Mentored and trained junior engineers on best practices in data engineering and AI/ML integration.
- Migrated legacy healthcare data systems to modern cloud-based data architectures, improving efficiency and reducing costs.
- Optimized NoSQL database queries and index design for performance improvements, resulting in a 30% reduction in query latency.
- Developed and optimized ETL pipelines using Azure Data Factory to transform and load data into the cloud environment, managing large datasets and ensuring data accuracy.
- Built custom data processing applications using Java to integrate external data sources, handle complex transformations, and load data into data warehouses.
- Integrated data processing workflows with Azure Fabric, ensuring smooth orchestration of complex, multi-step data workflows while improving reliability and performance.
- Designed and developed scalable ETL pipelines for data migration across multiple platforms (on-premises, cloud).
- Integrated Apache Kafka for real-time data streaming and used Docker containers to encapsulate microservices and ETL jobs, reducing operational overhead.
- Integrated containerized applications with OpenShift for enhanced application deployment and orchestration, improving release cycles and scalability.
- Spearhead the design, implementation, and optimization of end-to-end data pipelines using Azure Data Factory, enabling efficient data ingestion, transformation, and movement from on-premises and cloud sources to Azure SQL Database and Azure Synapse Analytics.
- Architected and implemented data lake solutions using Azure Data Lake Storage Gen2, optimizing storage capacity and ensuring high availability and cost-efficiency.

- Designed, developed, and optimized ETL pipelines using Azure Data Factory (ADF) for data ingestion from on-premises systems, SaaS platforms, and external APIs, integrating them into Azure Data Lake Storage for long-term storage and analytics.
- Assisted with Tableau dashboard creation, including report generation and performance tuning
- Designed and implemented a cloud-based data pipeline architecture using AWS services (S3, Lambda, Glue, Athena, Redshift) and Snowflake as the data warehousing solution.
- Created AWS Glue crawlers for crawling the source data in S3 and RDS.
- Collaborated with data scientists to optimize machine learning model deployment on GKE, resulting in faster model training and inference times.
- Lead the design, development, and deployment of big data solutions on Azure Databricks, enabling seamless data processing and analytics workflows.
- Used AWS Glue for transformations and AWS Lambda to automate the process.
- Architected and optimized ETL pipelines for structured and unstructured data using Azure Data Factory and Azure Databricks, improving data ingestion and transformation speed by 30%.
- Built and optimized ETL/ELT workflows using Azure Data Factory to streamline data ingestion from diverse sources.
- Led the design, development, and optimization of large-scale data integration and ETL pipelines using Talend.
- Orchestrated cloud-native infrastructure using Terraform and Kubernetes, ensuring cost efficiency and security best practices in cloud resource management.
- Led data migration initiatives from legacy systems to modern data platforms, achieving 99.9% data integrity and seamless transitions.
- Developed data models in Azure Synapse Analytics to support advanced healthcare reporting and insights.
- Ensured compliance with HIPAA and other healthcare data privacy regulations through robust security and governance frameworks.
- Architected and maintained a centralized data lake to consolidate structured and unstructured healthcare data.
- Created real-time data streaming pipelines using GCP Pub/Sub and Dataflow for financial data.
- Experience in GCP Dataproc, GCS, Cloud functions, BigQuery and moving data between GCP and Azure using Azure Data Factory.
- Collaborated with data scientists to deploy machine learning models for predictive analytics, including patient outcomes and risk assessments.
- Integrated data from Electronic Health Records (EHR) and other clinical systems into unified analytics platforms.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery
- Designed and implemented Power BI dashboards for actionable insights into healthcare KPIs, reducing reporting time.
- Leveraged Apache Spark and Databricks to process large datasets (petabytes of data) and run complex machine learning models, increasing processing efficiency by 40%.
- Conducted performance tuning and optimization for large-scale data processing in Databricks and Spark.
- Drove data quality initiatives by defining data validation and testing strategies within DBT models.
- Worked extensively with Azure Blob Storage and Azure Data Lake Storage to architect cloud-based storage solutions for raw, structured, and unstructured data.
- Architected and deployed a data warehouse on AWS Redshift, ensuring high availability and seamless integration with Tableau for real-time reporting.
- Automated data validation and quality checks to ensure data accuracy for critical decision-making.
- Supported real-time data streaming solutions using Event Hubs and Azure Stream Analytics.
- Partnered with cross-functional teams to define business requirements and align them with technical solutions.
- Created role-based access controls and encryption mechanisms to protect sensitive patient data.
- Led the implementation of data versioning and lineage tracking for regulatory compliance and audit readiness.
- Defined data governance policies and standards to improve data reliability and usability across healthcare applications.
- Built data pipelines to support clinical trial analytics and regulatory reporting.
- Deployed scalable CI/CD pipelines using Azure DevOps for automated deployment of data solutions.
- Mentored junior engineers on best practices for data engineering and healthcare domain knowledge.

- Conducted root cause analysis and resolved complex data discrepancies in clinical and operational datasets.
- Established a robust DevOps pipeline using Docker, Jenkins, and Kubernetes, automating data pipeline deployments and monitoring.
- Designed and executed proofs of concept to evaluate emerging data technologies for healthcare applications.
- Designed automated data ingestion processes using Scala and Apache Kafka, supporting data flows from various sources into centralized data lakes and warehouses.
- Optimized the ETL pipeline using Apache NiFi to streamline data ingestion from various sources (logs, APIs, etc.) into MongoDB.
- Developed metadata-driven frameworks for dynamic and reusable data workflows.
- Integrated third-party healthcare analytics tools with internal data platforms for enhanced insights.
- Optimized performance and reliability of large-scale data processing workloads using Kubernetes and cloud services, achieving a 30% reduction in processing time.
- Developed scripts using Spark, for loading the data from Hive to GCP Cloud SQL at a faster rate.
- Collaborated with healthcare practitioners to identify key analytics opportunities and deliver tailored solutions.
- Established monitoring and alerting systems for proactive issue resolution in data pipelines.
- Worked on interoperability standards such as HL7, FHIR, and DICOM for seamless data exchange in healthcare systems.
- Conducted training sessions for stakeholders on leveraging data platforms and analytics tools.
- Documented technical designs, workflows, and best practices for knowledge sharing across teams.
- Promoted innovation by exploring and integrating AI/ML capabilities for advanced healthcare analytics.
- Supported the optimization of supply chain operations in healthcare by analyzing operational data.

Environment: Azure Data Factory, Azure Synapse Analytics, Azure Databricks, Azure Data Lake Storage, Azure Blob Storage, Power BI, Azure SQL Database, Active Directory, Terraform, ARM Templates, PowerShell, REST APIs, HL7, FHIR, DICOM, HIPAA Compliance Standards, and CI/CD Pipelines.

Walmart, Linden, NJ **Aug 20 – Aug 21**
Sr. Data Engineer
Responsibilities:

- Accumulate the EEIM Alarm data to the NoSQL database called Mongo DB and retrieve it from Mongo DB when necessary.
- Process the system logs using log stash tool and store to elastic search and create dashboard using Kibana.
- Supported multiple data migration projects, moving data from SQL Server, Oracle, and MySQL to cloud-based platforms.
- Regularly tune performance of Hive queries to improve data processing and retrieving
- Integrated containerized applications with OpenShift for enhanced application deployment and orchestration, improving release cycles and scalability.
- Led the transition of on-premise infrastructure to Kubernetes-based cloud environments, ensuring smooth and scalable data operations.
- Spearheaded the migration of on-premises ETL workflows to Talend Cloud, ensuring zero data loss and enhanced scalability.
- Mentored junior engineers, promoting best practices in coding, database management, and MongoDB performance tuning.
- Implemented DBT for data transformation and testing, automating the creation of data models in Redshift and BigQuery.
- Utilized Terraform to automate GCP resource provisioning, implementing reusable modules for consistent and repeatable infrastructure setups.
- Established a robust DevOps pipeline using Docker, Jenkins, and Kubernetes, automating data pipeline deployments and monitoring.
- Spearheaded the migration of on-premises data systems to AWS Cloud and utilized Terraform to automate provisioning of cloud infrastructure resources.
- Designed and built scalable data warehouse solutions using Azure Data Warehouse (Synapse), improving data query performance and reducing reporting time.
- Optimized and automated data workflows using AWS Step Functions and AWS Batch on Amazon EMR for large-scale data processing tasks.
- Designed and implemented ETL pipelines using Python and Scala, integrating data from multiple disparate sources into a unified system.

- Managed the architecture and design of an enterprise-wide Azure Data Lake Storage solution for a large-scale data processing pipeline.
- Used Azure Fabric for orchestrating distributed data pipelines, improving data flow reliability and system resiliency by 40%.
- Utilized Azure Data Factory to automate data ingestion from multiple sources into Azure Data Lake Storage Gen2, improving data flow and accessibility.
- Developed Java-based data processing engines and backend services to automate the collection, transformation, and loading of data from relational and NoSQL databases.
- Led the integration of cloud-based ETL pipelines using Azure Data Factory, transforming raw data into actionable insights.
- Worked on integrating data from diverse sources, including relational databases, cloud platforms, and NoSQL systems, using Scala and Spark.
- Utilized Tableau for visualizing trends and KPIs, creating interactive dashboards that were used by management for operational decision-making.
- Managed and optimized clusters in GCP Dataproc and implemented resource allocation strategies using YARN based on the workflow demands to enhance processing efficiency.
- Create data ingestion modules using AWS Glue for loading data in various layers in S3 and reporting using Athena and Quick Sight.
- Performed Data Aggregation, Validation and on Azure HDInsight using spark scripts written in Python.
- Implemented Apache Spark-based processing for large datasets using Hadoop and Databricks, resulting in reduced job execution time.
- Performed monitoring and management of the Hadoop cluster by using Azure HDInsight.
- Development of Informatica Mappings, Sessions, Work lets, Workflows.
- Capable of using AWS utilities such as EMR, S3 and Cloud Watch to run and monitor Hadoop and Spark jobs on AWS.
- Developed performance tuning strategies for ETL jobs and managed resource allocation in Kubernetes clusters to optimize pipeline throughput.
- Developed efficient solutions for incremental data loads and real-time data replication to minimize data downtime during migrations.
- Used Oozie and Oozie Coordinators for automating and scheduling our data pipelines.
- Used AWS Athena extensively to ingest structured data from S3 into other systems such as Redshift or to produce reports.
- 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) and processing the data in In Azure Data bricks. Experience in Moving Data in and out of Windows Azure SQL Databases and Blob Storage.
- The Spark-Streaming APIs were used to conduct on-the-fly transformations and actions for creating the common learner data model, which receives data from Kinesis in near real time.
- Led the migration of on-premise data warehousing systems to Amazon Redshift and later to GCP BigQuery for improved scalability and performance.
- Implemented data ingestion from various source systems using Sqoop and Pyspark.
- Integrated Apache Kafka for real-time data streaming and used Docker containers to encapsulate microservices and ETL jobs, reducing operational overhead.
- Hands on experience implementing Spark and Hive jobs performance tuning.
- Develop framework for converting existing Power Center mappings and to Pyspark (Python and Spark) Jobs.
- Create Pyspark frame to bring data from DB2 to Amazon S3.
- Applied efficient and scalable data transformations on the ingested data using Spark framework.
- Created basic ETL workflows using SSIS and PowerShell for data extraction from legacy systems and loading into data warehouses.
- Provided seamless connectivity between BI tools like Tableau and Qlik to Redshift endpoints.
- Drove data quality initiatives by defining data validation and testing strategies within DBT models.
- Manage IAM roles and console access for EC2, RDS and ELB services.
- Worked on Big data on AWS cloud services i.e., EC2, S3, EMR and Dynamo DB
- Built S3 buckets and managed policies for it and used S3 bucket and Glacier for storage and backup on AWS.
- Along with Continuous Integration and Continuous Deployment with AWS Lambda and AWS code pipeline.

- Collaborated with Data Scientists and analysts to implement scalable machine learning pipelines using Azure Databricks and MLflow for model training, testing, and deployment.
- Analyzed data & defined KPIs; created tasks and set dependencies using Qlik View Publisher.
- Involved in performance tuning of various Qlik View applications.
- Created ad-hoc reports in Qlik View and supported the users with the ad hoc reporting queries.
- Provided demo of Qlik View dashboards to enhance user's knowledge on key capabilities of Qlik View application.
- Developed code to handle exceptions and push the code into the exception Kafka topic.
- Was responsible for ETL and data validation using SQL Server Integration Services.

Environment: PySpark, Map Reduce, HDFS, Azure Data Factory, Azure Synapse Analytics, Azure Databricks, Azure Data Lake Storage, Sqoop, flume, Kafka, Hive, Pig, HBase, SQL, Shell Scripting, Eclipse, SQL Developer, Git, SVN, JIRA, Unix.

Fairfax Financial, Miami, FL **Oct 18 – Jul 20**
Sr. Data Engineer
Responsibilities:

- Experience in Creating, developing, and deploying high-performance ETL pipelines with Pyspark and Azure Data Factory.
- Developed ETL pipelines in and out of data warehouse using a combination of Python, and Snowflake. Used Snow SQL to write SQL queries against Snowflake.
- Developed ELT processes from the files from abinitio, google sheets in GCP with compute being dataproc (pyspark) and big query.
- Implemented a 'serverless' architecture using API Gateway, Lambda, and Dynamo DB and deployed AWS Lambda code from Amazon S3 buckets.
- Supported multiple data migration projects, moving data from SQL Server, Oracle, and MySQL to cloud-based platforms.
- Designed and implemented data models for MongoDB collections, ensuring fast, reliable data retrieval for both real-time and batch processing scenarios.
- Designed and implemented a NoSQL-based solution (Cassandra) for a real-time recommendation engine, improving customer engagement by 25%.
- Managed ETL job scheduling and monitoring using Informatica PowerCenter and Informatica Cloud Monitoring tools.
- Improved data pipeline efficiency by optimizing Talend jobs and leveraging parallel processing and partitioning techniques.
- Designed automated data ingestion processes using Scala and Apache Kafka, supporting data flows from various sources into centralized data lakes and warehouses.
- Architected and deployed cloud-native data pipelines in AWS, leveraging Docker for containerization and Kubernetes for orchestration to ensure rapid deployment and high availability.
- Conducted training sessions on Kubernetes and containerization for the broader engineering team, fostering a DevOps culture.
- Designed and built real-time data pipelines to ingest and analyze streaming data using Azure Databricks and Azure Event Hubs.
- Developed and maintained DBT models, documentation, and tests to ensure consistent and accurate data transformations.
- Extensive use of cloud shell SDK in GCP to configure/deploy the services using GCP Big Query.
- Led the design and implementation of an Azure Data Lake to centralize financial data from various sources. Implemented automated ETL processes using Azure Data Factory and Scala to ensure smooth data flow for further analysis.
- Integrated Databricks with AWS Lambda for serverless data processing, enabling faster data ingestion and processing.
- Automated the entire cloud infrastructure deployment using Terraform, enabling version-controlled infrastructure as code (IaC) for multiple projects.
- Developed custom Python scripts for data processing and transformations, integrated with Azure Data Lake for scalable processing and storage.
- Enabled seamless integration of on-premises databases with Azure Data Lake using hybrid cloud capabilities, ensuring smooth data migration and storage.
- Worked on the migration of legacy data warehouses to Azure SQL Database and Azure Data Warehouse.
- Developed and deployed data pipelines on AWS Redshift and Snowflake for real-time and batch processing of large-scale datasets

- Architected and deployed a data warehouse on AWS Redshift, ensuring high availability and seamless integration with Tableau for real-time reporting.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
- Worked on an Azure copy to load data from an on-premises SQL server to an Azure SQL Data warehouse.
- Worked on redesigning the existing architecture and implementing it on Azure SQL.
- Experience with Azure SQL database configuration and tuning automation, vulnerability assessment, auditing, and threat detection.
- Integration of data storage solutions in spark - especially with Azure Data Lake storage and Blob snowflake storage.
- Involved in Migrating Objects from Teradata to Snowflake and created Snow pipe for continuous data load.
- Utilized Tableau for visualizing trends and KPIs, creating interactive dashboards that were used by management for operational decision-making.
- Assisted in building and maintaining a data warehouse environment using SQL Server and ETL pipelines.
- Improving the performance of Hive and Spark tasks.
- Knowledge with Kimball data modeling and dimensional modeling techniques.
- Created large datasets by combining individual datasets using various inner and outer joins in SAS/SQL and dataset sorting and merging techniques using SAS/Base.
- Extensively worked on Shell scripts for running SAS programs in batch mode on UNIX.
- Responsible for Building Cloud Formation templates for SNS, SQS, Elastic search, Dynamo DB, Lambda, EC2, VPC, RDS, S3, IAM, Cloud Watch services implementation and integrated with Service Catalog.
- Wrote Python scripts to parse XML documents and load the data in database.
- Used Hive, Impala and Sqoop utilities and Oozie workflows for data extraction and data loading.
- Created HBase tables to store various data formats of data coming from different sources.
- Responsible for importing log files from various sources into HDFS using Flume.
- Involved in Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Data bricks.
- Designed and managed Kubernetes clusters on AWS, using Helm for deployments and Kubernetes networking strategies to ensure high availability and fault tolerance.
- Designed and developed ETL pipelines to migrate transactional data from on-prem systems to data lakes in AWS and Azure.
- Designed SSIS Packages to transfer data from flat files, Excel SQL Server using Business Intelligence Development Studio.
- Developed a POC for project migration from on premise Hadoop MapR system to GCP.
- Analyzed the SQL scripts and designed the solution to implement using Pyspark.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query.
- Populated HDFS and HBase with huge amounts of data using Apache Kafka.
- Worked on Snowflake environment to remove redundancy and load real time data from various data sources into HDFS using Spark.
- Developed the DDLs for Hive and GCP Big Query and validation against source and target using complex SQL Queries.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
- Creating Data bricks notebooks using SQL, Python and automated notebooks using jobs.
- Creating Spark clusters and configuring high concurrency clusters using Azure Data bricks to speed up the preparation of high-quality data.
- Managed and optimized clusters in GCP Dataproc and implemented resource allocation strategies using YARN based on the workflow demands to enhance processing efficiency.
- 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.
- Installed Oozie workflow engine to run multiple Hive and Pig jobs which run independently with time and data availability.
- Involved in implementing security on Horton works Hadoop Cluster using with Kerberos by working along with operations team to move none secured cluster to secured cluster.
- Experienced with event-driven and scheduled AWS Lambda functions to trigger various AWS resources.

- Implemented and Developing Hive Bucketing and Partitioning.
- Implemented Kafka, spark structured streaming for real time data ingestion.

Environment: ADF, Data bricks and ADL Spark, Hive, HBase, Sqoop, Flume, ADF, Blob, cosmos DB, Map Reduce, HDFS, Cloud era, SQL, Apache Kafka, Azure, Python, power BI, Unix, SQL Server.

PNC Bank, Austin, TX **Aug 15 – Sep 18**
Data Engineer
Responsibilities:

- Leading efforts for migrating legacy systems to Microsoft Azure cloud-based solutions.
- Redesigning legacy application solutions to run efficiently on cloud platforms with minimal changes.
- Architecting data pipelines using Azure services like Data Factory to migrate data from legacy SQL servers to Azure databases.
- Led the migration of on-premises databases and data warehousing solutions to Azure Cloud infrastructure, ensuring high availability and scalability.
- Implementing API gateway services, SSIS packages, Talend jobs, and custom .NET and Python codes for data migration tasks.
- Optimized queries and data transformations in NoSQL databases (Couchbase, MongoDB), enhancing the overall data retrieval speed by 40%.
- Collaborated with data scientists to design feature engineering pipelines and preprocess data for machine learning models using MongoDB as the primary data store.
- Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift
- Designed and implemented data transformation logic using Talend components like tMap, tJoin, and tAggregateRow for complex data processing tasks.
- Spearheaded the migration of legacy data systems to Kubernetes, leveraging Docker containers and microservices architecture to improve deployment speed and scalability.
- Building pipelines to integrate Azure Cloud with AWS S3 to facilitate data transfer into Azure databases.
- Setting up Hadoop and Spark clusters for Proof of Concepts (POCs) and real-time streaming, integrating with ecosystems like Hive, HBase, and HDFS/Data Lake/Blob Storage.
- Enhanced data pipeline performance by tuning SQL queries, optimizing DBT models, and utilizing indexing and partitioning strategies.
- Configuring Spark clusters to process large volumes of data and transfer it to SQL Server, along with creating Spark jobs for data transformations and actions.
- Developing APIs to connect with media data feeds like Prisma, Double Click Management, Twitter, Facebook, Instagram, and Amnet, integrating with Cosmos DB.
- Implementing trigger-based mechanisms to optimize resource costs, such as Web Jobs and Data Factories, using Azure Logic Apps and Functions.
- Working extensively with relational databases, including Postgres SQL, as well as MPP databases like Redshift.
- Managed data migration projects from traditional on-premise databases to AWS and Snowflake, ensuring seamless data transfer with minimal downtime.
- Designed and optimized ETL workflows using AWS Glue and Lambda, automating data transformation and loading into data lakes and data warehouses.
- Designing custom processes for transformation using Azure Data Factory and automation pipelines.
- Leveraging Azure Data Factory and Logic App extensively for ETL processes, facilitating data movement between databases, Blob storage, HDInsight-HDFS, and Hive Tables.
- Collaborating with stakeholders to gather requirements and understand business needs for data architecture and application design.
- Designing and implementing scalable and efficient data architectures on Azure, considering performance, scalability, and cost-effectiveness.
- Developing and maintaining data pipelines for data ingestion, transformation, and loading using Azure Data Factory and other Azure services.
- Ensuring data quality, reliability, and consistency throughout the data pipeline by implementing data validation and error handling mechanisms.
- Monitoring and optimizing the performance of data pipelines and systems to ensure efficient data processing and resource utilization.
- Implementing security measures and access controls to protect sensitive data and ensure compliance with regulatory requirements.
- Collaborated with data scientists, analysts, and business stakeholders to gather requirements and create customized Tableau dashboards, providing critical insights to drive business decisions.

- Automating routine tasks and processes using Azure Automation and other automation tools to improve operational efficiency.
- Collaborating with cross-functional teams including developers, data scientists, and business analysts to integrate data solutions into applications and business processes.
- Documenting data architectures, processes, and workflows to ensure clarity and maintainability of data solutions.
- Providing technical guidance and mentoring to junior team members to facilitate their professional development.
- Staying updated with the latest Azure technologies and best practices in data engineering to drive innovation and continuous improvement.
- Strong experience in working with ELASTIC MAPREDUCE(EMR) and setting up environments on Amazon AWS EC2 instances
- Leveraged Docker and Kubernetes for containerization of data services, reducing operational overhead and improving system reliability.
- Conducting performance tuning and optimization of databases, data warehouses, and data processing systems to enhance performance and scalability.
- Troubleshooting data-related issues and providing timely resolution to ensure smooth operation of data solutions.
- Implementing disaster recovery and backup strategies to ensure data availability and resilience.
- Collaborating with cloud architects and infrastructure teams to ensure proper integration and deployment of data solutions within the Azure environment.
- Conducting code reviews, testing, and validation of data engineering solutions to ensure quality and reliability.

Environment: Azure Data Factory (ADF v2), Azure SQL Database, Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, ADLS Gen 2, Azure Cosmos DB, Azure Event Hub, Azure Machine Learning.

Cencora, Addison, TX, **Jun 11 – Jul 15**
Data Engineer
Responsibilities:

- Highly Involved into Data Architecture and ApplicaCon Design using Cloud and Big Data soluCons on AWS, Microsob
- Azure.
- Leading the effort for migraCon of Legacy-system to Microsob Azure cloud-based soluCon. Re-designing the Legacy
- ApplicaCon soluCons with minimal changes to run on cloud plaqorm.
- Built the data pipeline using Azure Service like Data Factory to load the data from Legacy SQL server to Azure Data
- Led the transition of on-premise infrastructure to Kubernetes-based cloud environments, ensuring smooth and scalable data operations
- Utilized AWS CloudWatch and X-Ray for monitoring Lambda performance, ensuring minimal latency and error handling for mission-critical workloads.
- Base using Data Factories, API Gateway Services, SSIS Packages, Talend Jobs, custom .Net and Python codes.
- 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 FuncCons.
- Build various pipeline to integrate the Azure Cloud to AWS S3 to get the data into Azure Database.
- Set up the Hadoop and Spark cluster for the various POCs, specifically to load the Cookie level data and real-Cme
- Assisted with the migration of large datasets from on-premise databases to AWS, leveraging AWS Data Pipeline and DynamoDB for data storage and manipulation.
- Developed and optimized ETL processes using AWS Glue and Snowflake's Snowpipe for continuous data ingestion.
- streaming. Integrate with other ecosystems like Hive, HBase, Spark, HDFS/DataLake/Blob Storage.
- Set up the Spark Cluster to process the more than 2 Tb of data and dumped into SQL Server. In addiCon, built
- Developed performance tuning strategies for ETL jobs and managed resource allocation in Kubernetes clusters to optimize pipeline throughput.
- various Spark jobs to run Data TransformaCons and AcCons.

- WriCng a different APIs to connect with the different Media Data feeds like, Prisma, Double Click Management,
- TwiNer, Facebook, Instagram and Amnet to get the Data using Azure Web Job and FuncCons integraCon with
- Cosmos DB.
- Built the trigger-based Mechanism to reduce the cost of different resources like Web Job and Data Factories using
- Azure Logic Apps and FuncCons.
- Extensively worked on RelaConal Database, Postgres SQL as well as MPP database like Redshib.
- Experience in Custom Process design of TransformaCon via Azure Data Factory & AutomaCon Pipelines.
- Extensively used the Azure Service like Azure Data Factory and Logic App for ETL, to push in/out the data from DB to
- Blob storage, HDInsight - HDFS, Hive Tables.

Environment: Azure Data Factory (ADF v2), Azure SQL Database, Azure functions Apps, Azure Data Lake, BLOB Storage, SQL server, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell