**Vikas**

**Senior Data Engineer**

**louis@cognitek.io**

**Phone: (732) 347 6970**

**Professional Summary:**

* 10+ years of experience as Data Engineer with strong expertise in design, development and maintenance of enterprise analytical solutions using Big Data Technologies.
* Solid grasp of Data Modelling, Data Validation, and Data Source Evaluation, along with a deep understanding of Data Warehouse and Data Mart Design, ETL, BI, OLAP, and Client/Server applications on AWS and Azure, Snowflake & MangoDB.
* platforms.
* Skilled in implementing CI/CD pipelines using tools like Jenkins, Git, and scripting in Linux for automation of ETL workflows and infrastructure provisioning
* Experience in utilizing analytical applications like Python to identify trends and relationships between different pieces of data, draw appropriate conclusions and translate analytical findings into risk management and marketing strategies that drive value.
* Proficient in data extraction, transformation, and loading using Azure Data Factory, T-SQL, and Spark SQL.
* Expertise in Big Data technologies including Hadoop, Spark, Python, SQL, Power BI, Tableau, and other Data Engineering tools.
* Strong experience with MongoDB for schema design, aggregation pipelines, data replication, indexing strategies, and integrating NoSQL data into analytical pipelines.
* Hands-on experience with Test-Driven Development (TDD) practices for ensuring reliable and maintainable code in Python, Java, and Scala.
* In-depth knowledge and implementation of ETL/ELT pipelines using Azure Data Factory, Snowflake, and Spark SQL, combined with strong Linux shell scripting capabilities for task automation.
* Specialization in data ingestion to Azure services such as Azure Data Lake, Azure storage, and SQL, followed by data processing in Azure Databricks and Data Factory.
* Hands-on experience with various Azure Cloud Technologies including Azure SQL Database, Data Factory, Azure Databricks, Azure Data Lake store, Azure Data Lake Analytics, Azure Synapse Analytics, Azure Key Vault.
* Hands on experience in all phases of SDLC mainly on Agile (Scrum) and Waterfall methodologies.
* Experience on SaaS, PaaS and IaaS concepts of Cloud Computing and Implementation.
* Experience working with Google Cloud (GCP) Services like Compute Engine, Cloud Functions, Cloud DNS, Cloud Storage and Cloud Deployment Manager.
* Expertise in data extraction, transformation, and loading using Azure Data Factory, T-SQL, and Spark SQL, combined with strong Linux shell scripting capabilities for task automation.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Experience in developing data pipelines for moving data from on-premises (Oracle) to Google Cloud Platform, transforming data from in Google Big Query.
* Experience working with AWS services like Amazon Redshift, Amazon RDS (equivalent to Cloud SQL), Amazon EMR (equivalent to Dataproc), Amazon S3, and Amazon DynamoDB.
* Extensive experience with AWS Big Data Services, including Redshift, EMR, and S3 for enterprise-level data modeling and implementation, ensuring efficient data management.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Good experience in setting up Python REST API Framework using Django.
* Experience in using stackdriver service/ dataproc clusters in GCP for accessing logs for debugging.
* Proficient of AWS services like VPC, Glue Pipelines, Glue Crawler, Cloud front, EC2, ECS, EKS, Elastic beanstalk, Lambda, S3 Storage, RDBS, Dynamo db, Redshift, Elastic Cache, DMS, SMS, Data Pipeline, IAM, WAF, Artifacts, API gateway, SNS, SQS, SES, Auto Scaling, Cloud Formation, Cloud Watch and Cloud Trail….
* Experience in handling operations and maintenance support for AWS cloud resources, including launching, maintaining, and troubleshooting EC2 instances, S3 buckets, Auto Scaling, DynamoDB, AWS IAM, and Elastic Load Balancers (ELB) and Relational Database Services (RDS). Also created snapshots for data to store in AWS S3.
* Strong expertise in using Hadoop big data technologies such as Apache Spark, Scala, Kafka, HDFS, Hive, Pig, MapReduce, Zookeeper, Sqoop, Oozie, Nifi, and Impala.
* Experience in setting up workflow using Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop jobs.
* Expertise in SQL Server Integration Services (SSIS) and SQL Server Reporting Services (SSRS) with good knowledge on SQL Server Analysis Services (SSAS).
* Experience working with various transformations like Normalizer, expression, rank, filter, group, aggregator, lookups, joiner, sequence generator, sorter, SQLT, stored procedure, Update strategy, Source Qualifier, Transaction Control, JAVA, Union, CDC etc.
* Strong background in version control systems like Git and GitHub for managing source code, collaborating on development efforts, and maintaining deployment pipelines.
* Expertise in using Python with OpenStack, OpenERP (now ODOO), SQLAlchemy, Django CMS.
* Excellent experience with Requests, NumPy, Matplotlib, SciPy, PySpark and Pandas python libraries during development lifecycle and experience in developing APIs for the application using Python, Django.
* Good experience on Data Modeling (Dimensional and Relational) concepts like Star-Schema Modeling, Snowflake Schema Modeling, Fact and Dimension Tables.
* Experience in developing custom-built ETL solution, batch processing, and real-time data ingestion pipeline to move data in and out of the Hadoop cluster using PySpark and Shell Scripting.
* Expertise in using ETL Tool Informatica Power Center 10.x/9.x/8.x (Designer, Workflow Manager, Repository Manager, ETL and Data Warehouse.
* Proven ability to design and implement end-to-end data pipelines across multi-cloud environments, handling batch and real-time ingestion, transformation, validation, and analytics using tools such as Databricks, Airflow, and Dataflow.
* Extensive experience with Jenkins for automating build, test, and deployment stages of data pipeline and application development projects.
* Experienced in using advanced concepts of Informatica like push down optimization (PDO).
* Experience in Performance Tuning and Debugging of existing ETL processes.
* Provided full life cycle support to logical/physical database design, schema management and deployment. Adept at database deployment phase with strict configuration management and controlled coordination with different teams.
* Strong working experience on NoSQL databases and their integration with the Hadoop cluster - HBase, Cassandra, MongoDB and DynamoDB.

**Technical Skills:**

|  |  |
| --- | --- |
| **Languages** | Python, Scala, Java, SQL, PL/SQL, PySpark |
| **Databases** | Oracle 9i/10g/11g/12, SQL Server 2000/2005, MS SQL, HBase, MongoDB, MySQL, Cassandra, Snowflake, DynamoDB, PostgreSQL |
| **Bigdata Technologies** | Apache Spark, Scala, Kafka, HDFS, Hive, Pig, MapReduce, Zookeeper, Sqoop, Oozie, Nifi, and Impala |
| **ETL Tools** | Informatica Power Center 10.x/9.x/8.x |
| **Development IDE`s** | Eclipse, Visual Studio Code, Toad, SQL Developer |
| **Logging & Monitoring** | Splunk, CloudWatch, Log4J, SLF4J, Zipkins, Grahana |
| **Operating Systems** | UNIX, Linux, Ubuntu, Windows XP/2000/VISTA |
| **Cloud Technologies** | AWS (Lambda, EC2, S3, SNS, CloudWatch, CloudFormation, RDS, VPC, Auto Scaling, IAM, AWS Glue, AWS Batch, AWS DMS, Code Build, Code Deploy), Microsoft Azure (Azure Databricks, Azure Data Factory, Azure Data Explorer, Azure HDInsight, ADLS), Google Cloud Platform (Big Query, Compute Engine, Cloud Functions, Cloud DNS, Cloud Storage, Cloud Deployment Manager). |
| **Version Control Tools** | CVS, SVN, GitHub, JIRA and Bitbucket |
| **Test Frameworks** | Junit, Mockito |
| **Methodologies** | Agile (Scrum), Waterfall |

**Professional Experience:**

**Client: Nationwide ,** **Ohio, United States Jan 2024 - Present**

**Role: Senior Data Engineer/Sr. Data Modeler/Sr. Business Data Analyst/Data ArchitectSr. Data Modeler/Sr. Business Data Analyst/Data Architect**

**Responsibilities:**

* Worked on multi cloud environment both AWS and GCP services.
* Designed and developed batch and streaming pipelines using AWS and GCP services for different clients.
* Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Apache PySpark, with Cloudera Distribution.
* Worked extensively on SQL, PL/SQL, Scala and UNIX shell scripting.
* Involved in Data mapping specifications to create and execute detailed system test plans. data mapping specifies what data will be extracted from an internal data warehouse, transformed, and sent to an external entity.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP.
* Developed ELT pipelines to ingest data from on-premise Oracle and flat files to Snowflake, using Streams, Tasks, and stored procedures for CDC and transformation logic.
* Experience in GCP Dataproc, Dataflow, Pub Sub, GCS, Cloud functions, Big Query, Stack driver, Cloud logging, Data studio for reporting.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Experience in GCP Data proc, GCS, Cloud functions, Big Query.
* Migrating an entire oracle database to Big Query and using power bi for reporting.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query
* Hands of experience in GCP, Big Query, GCS bucket and Stack driver.
* Used MongoDB as a data source for importing a part of client data.
* Write Scala program for spark transformation in Dataproc.
* Designed data solutions that integrated MongoDB as a NoSQL source into enterprise BI pipelines and unified views in Azure Synapse and Snowflake.
* Worked on POC to check various cloud offerings including Google Cloud Platform (GCP).
* Developed a POC for project migration from on prem Hadoop MapR system to GCP.
* Compared Self hosted Hadoop with respect to GCPs Data Proc, and explored Big Table (managed HBase) use cases, performance evolution.
* Develop and deploy the outcome using spark and scala code in Hadoop cluster running on GCP.
* Developed Python scripts to parse the Flat Files, CSV, XML, Scala, Terraform, JSON files and extract the data from various sources and load the data into data warehouse
* Good knowledge in building data pipelines in airflow as a service (composer) using various operators. Build a program using Python and Apache beam to execute it in cloud Dataflow and to run Data validation jobs between raw source file and big query tables.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big query.
* Extensive use of cloud shell SDK in GCP to configure/deploy the services using GCP BigQuery.
* Implemented data import and export from various sources through scripts and Sqoop, ensuring seamless data integration into the Big Data ecosystem.
* Extensive experience in designing, implementing, and managing Big Data solutions on the AWS cloud platform.
* Proficient in using core AWS services such as Amazon S3, EC2, EMR, Glue, Athena, Redshift, Lambda, and CloudFormation.
* Used Python, Scala, and Spark SQL to perform transformations and build reusable modules for cleansing, validation, and enrichment logic.
* Hands-on expertise in building scalable and reliable data pipelines using AWS services like AWS Glue, AWS Data Pipeline, or Apache Airflow.
* Developed custom UDFs in Python to extend the functionality of Hive and Pig Latin, enabling advanced data transformations and analytics.
* Migrated legacy data workflows from Hadoop (MapR) to GCP's BigQuery and Snowflake environment.
* Ability to design and implement scalable, fault-tolerant, and secure Big Data architectures on AWS using best practices such as well-architected frameworks and serverless computing.
* Collaborated with data stewards, SMEs, and business stakeholders to design optimal data models (star/snowflake schema) and ensure high-quality documentation.
* Proficient in using infrastructure-as-code tools like AWS CloudFormation or Terraform for automated provisioning and management of cloud resources.

**Environment:**Hadoop, Cloudera, MapReduce, Kafka, Impala, Spark, Zeppelin, Hue, NoSQL, Pig, Hive, Sqoop, GCP, Java, Scala, Cassandra, SQL, Tableau, ZooKeeper, Teradata, Zoom-Data, Linux Red-Hat, Oracle 12c.

**Client:** **Hewlett Packard Enterprise, Texas July 2020 – Dec 2023**

**Role: Senior Big Data Engineer**

**Responsibilities:**

* Worked in multi cloud architecture designing, building multiple Data pipelines, end to end ETL from Data ingestion and transformation in GCP and AWS data pipelines.
* Implemented Big Query data process from GCP Pub/Sub topic to Bigqueryusing cloud Dataflow with python and also used rest API with python to ingest Data from other systems into BigQuery.
* Executed cloud dataflow to run Data Validation between raw source file and Bigquery and integrated monitoring Bigquery, Dataproc using Stackdriver across environments.
* Wrote Perl scripts covering data feed handling, implementing business logic, communicating with web services through SOAPLite module and WSDL.
* Generated the data cubes using hive, Pig, JAVA Map-Reducing on provisioning Hadoop cluster in AWS.
* Created S3 buckets (configure, policies, permissions) and used AWS S3 for storage and backup of Data in to AWS and AWS Glacier to store archive data.
* Implemented AWS solutions using E2C, S3, RDS, EBS, Elastic Load Balancer, Glue Pipelines, Glue Crawler, Auto scaling groups, Optimized volumes, and EC2 instances and created monitors, alarms, and notifications for EC2 hosts using Cloud Watch.
* Implemented AWS Step Functions to automate and orchestrate the Amazon sagemaker related tasks such as publishing data to S3, training ML model and deploying it for prediction.
* Handled AWS EC2, Amazon S3, Amazon RDS, Elastic Load Balancer, Auto Scaling, CloudWatch, SNS, Amazon Lambda andStepFunctions.
* Generated consumer group lags from Kafka using their API Kafka. Used for building real-time data pipelines between clusters. Extracted files from Hadoop and dropped on daily hourly basis into S3.
* Authored Python (PySpark) Scripts for custom UDF's for Row/Column manipulations, merges, aggregations, stacking, data labelling and for all Cleaning and conforming tasks.
* Migrated an entire oracle database to BigQuery and build Data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Worked on Data Modeling involving Dimensional Data modeling, Star Schema/Snowflake schema, FACT & Dimensions tables, Physical & logical data modeling.
* Designed Snowflake Schema for Data Warehouse, ODS architecture by using tools like Data Model, Erwin.
* Developed data models and data migration strategies utilizing concepts of snowflake schema.
* Developed Kibana Dashboards based on the Log Stash data and integrated different source and target systems into ElasticSearch for near real time log analysis of monitoring End to End transactions.
* Worked on Informatica power center Mappings, Mapping Parameters, Workflows, Variables and Session Parameters.
* Wrote Pig Scripts to generate MapReduce jobs and performed ETL procedures on the data in HDFS.
* Developed solutions to leverage ETL tools and identify opportunities for process improvements using Informatica and Python.
* Participated in a Scrum team and Agile Practices including Test-Driven Development (TDD), Behaviour-Driven Development (BDD) and pair programming.
* Worked on various automation tools like GIT and Dimensional Modeling Star Schema, Snowflake schema, transactional modeling.
* Studied the existing system and conduct reviews to provide a unified review on jobs.
* Involved in testing the database using complex SQL scripts and handling the performance issues effectively.

**Environment:** Hadoop, Apache Spark, Scala, HDFS, Kafka, Hive, GitHub, Google Cloud Platform (GCP), Amazon Web Services (AWS), Python, PySpark, Jenkins, Perl, Agile, Informatica Power Center, UNIX, Snowflake

**Client:** **AbbVie** **North Chicago, Illinois Mar 2018 – June 2020**

**Role: Senior Data Engineer**

**Responsibilities:**

* Migrated an existing on-premises application to Amazon Web Services (AWS) and used its services like EC2 and S3 for small data sets processing and storage, experienced in maintaining the Hadoop cluster on AWS EMR.
* Developed solutions to pre-process large sets of structured, semi-structured data, with different file formats like Text, Avro, Sequence, XML, JSON, and Parquet.
* Developed an ETL process in AWS Glue to migrate customer data from external data stores such as S3 into AWS Redshift.
* Built pipelines to copy the data from multiple sources to destination in AWS Redshift.
* Migrated the data from Redshift data warehouse to Snowflake database.
* Built dimensional modelling, data vault architecture on Snowflake.
* Built scalable distributed Hadoop cluster running Hortonworks Data Platform (HDP)
* Developed Spark code using Scala and Spark-SQL for faster testing and processing of data and exploring of optimizing it using SparkContext, Spark-SQL, PairRDD's
* Serialized JSON data and storing the data into tables using Spark SQL
* Used Spark Streaming and collected data from Kafka in near-real-time and performs necessary transformations and aggregation to build the common learner data model and stores the data in NoSQL store (HBase).
* Worked on Spark framework on both batch and real-time data processing
* Worked in MLlib from Spark are used for predictive intelligence, customer segmentation and for smooth maintenance in Spark streaming.
* Implemented and maintained Hadoop cluster on AWS EMR.
* Loaded data into S3 buckets using AWS Glue and Spark.
* Implemented Spark in EMR for data processing in AWS Data Lake.
* Involved in designing and Developing Spark workflows using Scala to pull the data from the AWS S3 bucket.
* Utilized AWS Glue for data cataloging, ETL processing, and data preparation tasks, enabling seamless integration of diverse data sources and efficient transformation of large-scale datasets.
* Leveraged Glue's built-in connectors and transformations to handle complex data structures, such as nested data and JSON formats, facilitating effective data processing and integration for analytics and reporting purposes.
* Proficient in AWS Lambda for developing scalable and cost-effective applications.
* Implemented AWS Step Functions to orchestrate complex workflows in a serverless architecture.
* Utilized Amazon Redshift for high-performance data warehousing solutions.
* Optimized Redshift clusters through performance tuning, query optimization, and data distribution strategies.
* Utilized Amazon Cloud Watch for monitoring and gaining insights into application and infrastructure performance.
* Implemented AWS CI/CD pipelines using services like AWS Code Pipeline, AWS Code Build, and AWS Code Deploy.
* Developed programs for Spark streaming which takes the data from Kafka and pushes into different sources
* Loaded the data from the different Data sources like (Teradata, DB2, Oracle and flat files) into HDFS using Sqoop and load into Hive tables, which are partitioned.
* Created different pig scripts & converted them as shell command to provide aliases for common operation for project business flow. Implemented Partitioning, Bucketing in Hive for better organization of the data.
* Created few Hive UDF's to as well to hide or abstract complex repetitive rules.
* Converted Hive/SQL queries into Spark transformations using Spark DataFrames and Scala.
* Used different tools for data integration with different databases and Hadoop.
* Developed Oozie Workflows for daily incremental loads, which gets data from Teradata and then imported into hive tables.
* Developed bash scripts to bring log files from FTP server and then processing it to load into Hive tables.
* Developed a NiFi Workflow to pick up the data from Data Lake as well as from server and send that to Kafka broker
* Involved in loading and transforming large sets of structured data from router location to EDW using an Apache NiFi data pipeline flow.
* Implemented Kafka event log producer to produce the logs into Kafka topic which are utilized by ELK (Elastic Search, Log Stash, Kibana) stack to analyze the logs produced by the Hadoop cluster

**Environment:** Python, AWS, EC2, S3, EMR, Redshift, Hadoop, MapReduce, Hive, Pig, Spark, Kafka, Oozie, Nifi, Scala, PySpark, Snowflake, HBase, SQL

**Client: Black Knight, Jacksonville, FL**  **Feb 2017 – Feb 2018**

**Role: Data Engineer**

**Responsibilities:**

* Created Pipelines in ADF using Linked Services/Datasets/Pipeline to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, and write-back tool.
* Build a scheduling framework using python SDK to launch and collect status for all ADF jobs. The status is stored on the SQL database so that real-time dashboards could be built for visual representation of pipelines.
* Developed Python UDF for handling nested JSON data from the source system and flattening to line-item level records. These flattened records are further transformed using Spark transformations for daily aggregations and reporting.
* Developed data pipelines and ETL workflows in Azure Synapse Analytics using Azure Data Factory, Azure Databricks, and Azure Synapse Studio.
* Built and deployed Azure Synapse Analytics pipelines and workflows using Azure DevOps and Azure Resource Manager templates.
* Integrated Azure Synapse Analytics with other Azure services such as Azure Blob Storage, Azure Event Hubs, and Azure Key Vault.
* Implemented SELF-HOSTED integration runtime on Windows server to establish a secure connection between the Hadoop cluster and Azure data factory to migrate data from HDFS to Azure Data Lake.
* Implemented versioning through Azure GIT on Azure Data factory and scheduled all pipelines through Scheduled Triggers.
* Used Kafka and Spark Streaming for data ingestion and cluster handling in real-time processing. Developed flow XML files using Apache NIFI, a workflow automation tool to ingest data into HDFS.
* Involved in Designing Snowflake Schema for Data Warehouse, ODS architecture by using tools like Data Model, Erwin.

**Environment:** Azure Data Factory (ADF), Azure Dataflow, Apache NiFi , Azure Event Hubs, Azure Event Hubs, Azure Stream Analytics, Apache Kafka, Azure SQL, Blob storage, Azure SQL Data Warehouse, Azure Synapse Analytics, Azure Databricks, Snowflake Schema, Python SDK, Spark Structured Streaming, Azure DevOps, Azure Resource Manager templates, Azure GIT, Azure Key Vault, Erwin (Data Modeling tool)

**Client: Lazard , Hyderabad,** **India. June 2013 - Nov 2015**

**Role: ETL Developer**

**Responsibilities:**

* Developed Advance PL/SQL packages, procedures, triggers, functions, Indexes and Collections to implement business logic using SQL Navigator.
* Generated server-side PL/SQL scripts for data manipulation and validation and materialized views for remote instances.
* Created management analysis reporting using Parallel Queries, Java Stored Procedures. Participated in change and code reviews to understand the testing needs of the change components. Worked on troubleshooting defects in a timely manner.
* Involved in defragmentation of tables, partitioning, compressing and indexes for improved performance and efficiency. Involved in table redesigning with implementation of Partitions and Partition Indexes to make database performance and easier to maintain.
* Experience in Database Application Development, Query Optimization, Performance Tuning and DBA solutions and implementation experience in complete System Development Life Cycle.
* Used Informatica Power Center Designer to analyze the source data to Extract & Transform from various source systems by incorporating business rules using different objects and functions that the tool supports.
* Used Power Center Designer to create mappings and mapplets to transform the data according to the business rules.
* Used various transformations like Source Qualifier, Joiner, Lookup, SQL, Router, Filter, Expression and Update Strategy etc.
* Created and configured Workflows and Sessions to transport the data to target Oracle tables using Informatica Workflow Manager.
* Implemented complex business rules in Informatica Power Center by creating re-usable transformations, and robust Mapplets.
* Implemented performance tuning of Sources, Targets, Mappings and Sessions by identifying bottlenecks and used debuggers to debug the complex mappings and fix them.
* Designed and developed Informatica Workflow to extract data from XML files and loaded it into the database.
* Improved session Performance by enabling property incremental aggregation to load incremental data into the target table.
* Implemented performance tuning logic on Targets, Sources, Mappings and Sessions to provide maximum efficiency and performance.
* Worked along with the UNIX team for writing UNIX shell scripts to customize the server scheduling jobs.
* Used principles of Normalization to improve the performance. Involved in ETL code using PL/SQL in order to meet requirements for Extract, transformation, cleansing and loading of data from source to target data structures.
* Involved in the continuous enhancements and fixing of production problems. Designed, Implemented and Tuned interfaces and batch jobs using PL/SQL.
* Involved in data replication and high availability design scenarios with Oracle Streams. Developed UNIX Shell Scripts to automate repetitive database processes.
* Written documentation to describe program development, logic, coding, testing, changes and corrections.

**Environment:** Oracle 10g/11g, SQL Plus, TOAD, SQL Loader, SQL Developer, PL/SQL, Informatica Power Center, Designer, Workflow Manager, Workflow Monitor, Repository Manager, Shell Scripts, UNIX, Windows XP, Splunk, HTML, TOAD, XML.