**Deep Patel | Sr Data Engineer | Deeppateltestfour@gmail.com | +1 813 388 8735**

Certified Professional data engineer with over 9+ years of experience in various industries with various technologies. Able to provide smart, automated solutions in data monitoring, ETLs, and management with expertise in gathering requirements prioritizing tasks, and streamlining processes to minimize the need for manual intervention.

**Technical Summary**

* Experience with various cloud-based technologies like AWS, Azure and GCP.
* Expert in working with various Big Data Distributions like Cloudera, EMR serverless, HortonWorks and MapR.
* Experience in using various tools like Sqoop, Flume, Kafka, and Pig to ingest structured, semi-structured, and unstructured data into the cluster.
* Experience in working with Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and am familiar with building custom Airflow operators and orchestrating workflows with dependencies involving multi-clouds.
* Experience with Apache Spark and Spark Streaming using Scala and Python.
* Developed highly optimized Spark applications to perform various data cleansing, validation, transformation, and summarization activities according to the requirement.
* Designed and developed logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema and Slowly Changing Dimensions.
* Experience in Infrastructure Development and Operations involving AWS Cloud Services, EC2, EBS, VPC, RDS, SES, ELB, Auto scaling, CloudFront, Cloud Formation, Elastic Cache, API Gateway, Route 53, Cloud Watch, SNS.
* Experience in changing over existing AWS infrastructure to Serverless architecture (AWS Lambda, AWS Kinesis) through the creation of a Serverless Architecture using AWS Lambda, API gateway, Route 53, S3 buckets.
* Experience in implementing data models using HBase, Dynamo DB, and Cassandra.
* Expertise in building CI/CD on AWS environment using AWS Code Commit, Code Build, Code Deploy, and Code Pipeline and experience in using AWS CloudFormation, API Gateway, and AWS Lambda in automation and securing the infrastructure on AWS.
* Experience in implementing Data warehouse solutions in Confidential Redshift; Worked on various projects to migrate data from on-premises databases to Redshift, RDS, and S3.
* Experienced in processing real-time data using Kafka 0.10.1 producers and stream processors and implemented stream process using Kinesis and data landed into Datalake S3.
* Experience in extracting and loading data using complex business logic using Hive from different data sources and building ETL pipelines to process terabytes of data daily.
* Experience with implementing ETL transformations in Snowflake and Airflow.
* Implemented Security roles on SSAS cubes and Power BI services by implementing authentication and authorization methods.
* Experience in migrating projects from traditional RDBMS to Big Data platforms.
* Exposure in automating applications using CI and CD tools like Jenkins and Drone.
* Exposure to containerization tools like Docker and Kubernetes.
* Worked on integration tools like Talend, Apache NIFi for ingesting data from many sources.
* Experience with a messaging queue like Apache Kafka and Rabbit MQ.
* Experience on GCP tools like BigQuery, Pub/Sub, Cloud SQL, and Cloud functions.
* Experience with tensor flow libraries to build pipelines.
* Experience in building metrics dashboards and alerts using Grafana and Kibana.

**Professional Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **Lincoln Financials** | Location | **PA** |
| Designation | **Sr Cloud Engineer** | Duration | **July 2022- Current** |

**Responsibilities**

* Implemented various servers for production applications using Data lakes, and Dataproc.
* Import the data from different sources like HDFS/HBase into Spark RDD and develop a data pipeline using Kafka and Storm to store data into HDFS.
* Designed and developed Azure Cloud relational servers and databases analyzing current and future business requirements.
* Worked on migration of data from on Prem SQL server to Cloud database (Azure Synapse, Databricks, Analytics (DW) & Azure SQL DB).
* Created tabular models on Azure analysis services for meeting business reporting requirements.
* Used Azure services like ADLS and Synapse analytics for small data sets.
* Experience in data transformations using Azure HDInsight, HIVE for different file formats.
* Developed Spark and SparkSQL code to process the data in Apache Spark on Azure HDInsight to perform the necessary transformations based on the STMs developed.
* Worked on Version control systems like Subversion, Azure DevOps, GIT, CVS.
* Developed business intelligence solutions using SQL Server data tools to load data to SQL & Azure Cloud databases.
* Designed and implemented Unix-based solutions for data integration in collaboration with cross-functional teams, ensuring data consistency across various sources and formats.
* Proficient in designing, deploying, and managing data processing pipelines on Google Kubernetes Engine (GKE) to ensure scalable and containerized data applications.
* Used Airflow for scheduling the Hive, Spark and Map Reduce jobs.
* Developed Oozie work processes for planning and arranging the ETL cycle. Associated with composing Python scripts to computerize the way towards extricating weblogs utilizing Airflow DAGs.
* Transformed date related data into application compatible format by developing Apache Pig UDFs. Data Ingest from Sqoop & flume from Oracle database.
* Wrote the Pig UDF in for converting Date format and time stamp formats from the unstructured files to required date formats and processed the same.
* Hands on Abinitio ETL, Data Mapping, Transformation and Loading in complex and high-volume environment.
* Analyzed and managed system logs with Splunk and syslog.
* Assisted the Product Owner in creating a Proof of Concept by analyzing months of syslog data in the proposed Splunk platform in terms of reporting, monitoring and alerts.
* Imported/exported data into HDFS and Hive using Sqoop and Kafka.
* Experienced in developing Map Reduce programs using Apache Hadoop for working with Big Data.
* Validate Scoop jobs, Shell scripts & perform data validation to check if data is loaded correctly without any discrepancy. Perform migration and testing of static data and transaction data from one core system to another.
* For data extraction, transformation, and loading (ETL) tasks, developed and maintained intricate Unix shell scripts to guarantee accurate and efficient data flow.
* Proficient in deploying and managing data processing pipelines on Google Kubernetes Engine (GKE) for scalable and containerized data applications.
* Frameworks for distributed data processing, including Apache Spark and Apache Flink, have been successfully installed on GKE, allowing for effective parallel processing of large datasets.
* Applied Apache Kafka to transform live streaming with the batch processing to generate reports.
* Extensively using open source languages Perl, Python, Scala and Java.
* Writing scripts for creating, truncating, dropping, altering HBase tables to store the data after execution of MapReduce job and to use that for later analytics.
* Unix scripts have error handling and logging mechanisms in place to find and fix data anomalies and processing errors, ensuring the quality and integrity of the data.
* Create Self Service reporting in Azure Data Lake Store Gen2 using an ELT approach.
* Developed data warehouse model in snowflake for over 100 datasets using where cape.
* Experience in Agile Methodologies, Scrum stories and sprints experience in a Python based environment, along with data analytics and data wrangling.
* Performance tune up Phoenix/HBase, Hive queries and Spark.
* The custom File System plugin allows Hadoop Map Reduce programs, HBase, Pig and Hive to work unmodified and access files directly.
* Writing PySpark and Spark SQL transformation in Azure Data bricks to perform complex transformations for business rule implementation
* Design and develop hive, HBase data structure and Oozie workflow.
* Extending Hive and Pig core functionality by writing custom UDFs, UDTF and UDAFs.
* Build and maintain the environment on Azure IAAS, PAAS.
* GKE was used to containerize data processing applications, and the microservices architecture was used for increased scalability and flexibility.
* Implement Continuous integration/continuous development best practice using Azure DevOps, ensuring code versioning
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB).
* Extensive usage of Azure Portal, AWS omics, Aws glue, Azure PowerShell, Storage Accounts, Certificates and Azure Data Management.

**Environment:** GCP, BigQuery, Cloud SQL, AWS Batch, EMR serverless, NF-Core Apache Beam, Dataflow, Airflow, Pyspark, Python, Spark, Jupyter notebooks, Dataproc,Data Lakes, Databricks.

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **HCSC** | Location | **CA** |
| Designation | **Sr Data Engineer** | Duration | **June 2021 – July 2022** |

Responsbilities

* Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks, etc.
* Analyze escalated incidences within the Azure SQL database. Implemented test scripts to support test-driven development and continuous integration.
* Provided technical solutions on MS Azure HDInsight, Hive, MongoDB, Power BI, Spot Fire, Tableau, Azure SQL Data Warehouse Data Migration Techniques using BCP, Azure Data Factory.
* Data Migration from existing Teradata Systems to Hortonworks HDInsight cluster on Azure
* Used Azure Data Factory as an orchestration tool for integrating data from upstream Systems.
* Automated jobs using different triggers (Event, Scheduled, and Tumbling) in ADF.
* 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.
* Implemented Security roles on SSAS cubes and Power BI services by implementing authentication and authorization methods.
* Developed different reports using Microsoft BI tools like SSIS and SSRS and developed Power BI dashboards that would help businesses operate smoothly.
* Creating Spark clusters and configuring high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Responsible for ingesting data from various source systems (RDBMS, Flat files, Bigdata) into Azure (Blob Storage) using the framework model.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Expert working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Planning and design the solution to implement, using Azure Files Share.
* Propose architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.
* Worked on Azure Templates for quick deployments and advanced PowerShell scripting.
* Implemented HA with Azure Classic and Azure Resource Manager deployment models.
* Using the Azure SDK, automated the testing of data pipelines and implemented a data quality check framework in Python.
* Experienced in Gitlab CI and Jenkins for CI and End-to-End automation for all build and CD.
* Implemented automated alerting and monitoring systems on top of our applications, which are built using tools like Grafana and Kibana. So, we would receive automated alerts in case of any production failures to report any on-call rotation person in the team.
* Extensive experience with Azure Delta Lake, Azure Databricks, Azure Data Factory, Azure HD Insights, and Azure functions.
* Worked closely with data scientists for building predictive models using PySpark.
* Expertise in Architectural blueprints and detailed documentation. Create a bill of materials, including required cloud services and tools.

Environment: Azure Databricks, AWS Omics, Azure Datalake Gen 1, Azure data factory, EMR serverless, AWS Glue, Databricks Delta Lake, Snowflake, Spark, HIVE, Kafka, JMS, Oracle, Teradata, Python, Tableau, Power Bi

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **Doordash** | Location | **CA** |
| Designation | **Sr Data Engineer** | Duration | **Sep 2018 – May 2021** |

Responsbilities

* Worked in an Agile environment, and used the rally tool to maintain the user stories and tasks.
* Implemented Apache Sentry to restrict access to the hive tables on a group level.
* Designed and implemented by configuring Topics in the new Kafka cluster in all environments.
* Created multiple dashboards in Tableau for multiple business needs.
* Implemented Partitioning, Dynamic Partitions, and Buckets in HIVE for efficient data access.
* Designed SSIS Packages to extract, transfer, and load (ETL) existing data into SQL Server from different environments for the SSAS cubes (OLAP)
* Extract Transform and Load data from source systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, Azure Data Lake analytics.
* Designed & and implemented database solutions in Azure SQL Data Warehouse, Data Lake, and Azure SQL.
* Implemented Composite server for the data virtualization needs and created multiple views for restricted data access using a REST API.
* Vertical and horizontal scaling strategies were used to optimize the use of resources on GKE clusters, effectively controlling infrastructure costs while maintaining performance.
* Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team Using Tableau.
* Migrated Map reduce jobs to Spark jobs to achieve better performance.
* Involved in converting Map Reduce programs into Spark transformations using Spark RDD's using Scala and Python.
* Worked cross-functionally with software engineers and data scientists to design customized GKE solutions that satisfied particular data processing specifications, ensuring peak performance.
* Developed Apache Spark applications by using spark for data processing from various streaming sources.
* Developed data pipeline using Spark, Hive, Pig, python, Impala, and HBase to ingest customer
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Queried and analyzed data from Cassandra for quick searching, sorting and grouping through CQL.
* Joined various tables in Cassandra using spark and Scala and ran analytics on top of them.
* GKE configurations, deployment processes, and troubleshooting procedures were kept in depth documented.
* Applied Spark advanced procedures like text analytics and processing using the in-memory processing.
* Implemented Apache Drill on Hadoop to join data from SQL and No SQL databases and store it in Hadoop.
* Brought data from various sources in to Hadoop and Cassandra using Kafka.
* SQL Server reporting services (SSRS). Created & formatted Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Sub reports, ad-hoc reports, parameterized reports, interactive reports & custom reports.
* Created Kubernetes Deployments and Services to manage and expose containers, ensuring high availability and load balancing.
* Designing and Developing Oracle PL/SQL and Shell Scripts, Data Import/Export, Data Conversions and Data Cleansing

Environment: PySpark, Spark RDD, S3. JSON, Parquet, ETL, Hive, UNIX, HDFS, Power BI, SQL Server. DevOps, Kerberos, Scala, Python, AWS, Databricks, Kafka, Redshift, Airflow, Snowflake, Cassandra, Map Reduce (EMR), AWS Athena, Linux, Ubuntu, EC2, MongoDB, JavaScript, Ajax, CAT, AWS Glue, Crawlers, Cloudera, Jenkins. data bricks.

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **Western Union** | Location | **CO** |
| Designation | **Data Engineer** | Duration | **June 2017 – Aug 2018** |

Responsbilities

* Experience in developing SFTP, and NAS integrations to ingest data into HDFS using Python.
* Implemented messaging queues and routes using microservices camel in java.
* Developed batch ingestion jobs from Teradata to HDFS and Hive using Sqoop.
* Developed data lake platform using multiple tools like Kafka, Sqoop, Hive, Spark and Oozie.
* Implemented end-to-end job automation in Hadoop using Apache Oozie.
* Developed transnational system updates in HBase for data lake implementations.
* Developed end to end ETL operations in optimized way using Hive and Spark.
* Expert in handling complex data issues and memory optimizations and tuning in Spark.
* Implemented multiple data pipelines using Apache Spark using python and Scala.
* Developed real time streaming application to ingest Json messages using Apache Kafka.
* Implemented Data security features in data exposed through API endpoints.
* Expert in writing in many features using scripting languages like Bash, Shell and Python.
* Worked on implementing CRUD operations in HBase for multiple applications.
* Handling the Tickets/Service Calls raised by End-users and providing them faster resolution.
* Implemented multiple Change Request includes of new developments.
* Responsible for developing Python wrapper scripts which will extract specific date range using Sqoop by passing custom properties required for the workflow
* Involved in filtering data stored in S3 buckets using Elastic search and loaded data into Hive external tables.
* Designed and developed UDF'S to extend the functionality in both PIG and HIVE
* Import and Export of data using Sqoop between MySQL to HDFS on regular basis
* Developed a shell script to create staging, landing tables with the same schema as the source and generate the properties which are used by Oozie Jobs.
* Worked with NoSQL databases like Databricks, HBase in creating HBase tables to load large sets of semi structured data coming from various sources
* Developed Oozie workflows for executing Sqoop and Hive actions.
* Built various graphs for business decision making using Python matplotlib library

Environment: Hortonworks, Spark, Hive, Oozie, Databricks, Bash, Shell, Python, HBase, Teradata, Sqoop, Scala, data lake, SFTP, microservices, Java, Kafka, API, SQL.

|  |  |  |  |
| --- | --- | --- | --- |
| Client | **DCM Infotech** | Location | **India** |
| Designation | **Data Analyst** | Duration | **Aug 2015 - Dec 2016** |

Responsbilities

* Understand the data visualization requirements from the Business Users.
* Writing SQL queries to extract data from the Sales data marts as per the requirements.
* Developed Tableau data visualization using Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
* Tuning Oracle databases and tuning applications (SQL), for tuning SQL working closely with development Team.
* Developing containment scripts for data reconciliation using SQL and Python.
* Performed data analysis and data profiling using complex SQL on various sources systems including MySQL and Teradata.
* Evaluated data profiling, cleansing, integration, and extraction tools (e.g., Informatica)
* Wrote user defined functions (UDFs) in Hive to manipulate strings, dates, and other data.
* Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python.

Environment: Informatica, Data Analysis, Excel, SQL, Hive, UDFs, Python, Data Lake.