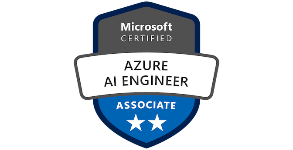
**Tarun Kumar D **

**Email:** [**tarunkumara444@gmail.com**](mailto:tarunkumara444@gmail.com)

**Contact: (940)-394-0399**

**Senior AI/ML & Cloud Data Engineer**

**Professional Summary:**

* Results-driven **Senior AI/ML & Cloud Data Engineer** with **9+ years of experience** designing and delivering scalable, cloud-native data and machine learning solutions across **AWS, Azure, and GCP**. Strong background in application development, data engineering, and AI/ML integration using **Python, Java, Scala, and SQL**.
* Extensive experience in **data ingestion, transformation, and orchestration** using **Apache Airflow**, with hands-on implementation of custom DAGs, operators, and workflow scheduling across multi-cloud environments.
* Proficient in building **enterprise-scale data pipelines** and **stream processing frameworks** using **Apache Kafka**, **Spark Streaming**, and **Google Dataflow**, ensuring real-time, fault-tolerant data processing.
* Designed and deployed **AI/ML workflows** using **Spark MLlib, TensorFlow, Azure ML, and GCP Vertex AI**, focusing on predictive analytics, classification, and **LLM-based applications** including **Retrieval-Augmented Generation (RAG)** and feature extraction.
* Experienced in message-based and file-based integrations, including secure, encrypted file processing and XML data transformation.
* Familiar with cloud-to-cloud and hybrid cloud (on-prem to cloud) data workflows, leveraging VPC, VPN, and private link setups
* Skilled in **data modeling and warehousing** using **Snowflake**, **BigQuery**, and traditional modeling techniques like **Star and Snowflake Schema**, driving scalable and performant analytics.
* Strong expertise with **Azure services** such as **Data Factory, Data Lake, Blob Storage, Synapse**, and **Databricks**, as well as **GCP services** including **BigQuery, Cloud Storage, Dataflow, Pub/Sub**, and **Cloud Functions**.
* Implemented **HIPAA-compliant** cloud architectures with a strong focus on **security best practices**, including IAM, VPC setup, key management (KMS), and encryption on **AWS, Azure, and GCP**.
* Experienced with **AWS analytics and compute services** such as **EC2, S3, EMR, Redshift, Lambda, Glue, CloudWatch**, and **Athena**, enabling reliable and cost-efficient cloud data workflows.
* Expert in **Spark development** using **Scala, Python, and Java**, building high-throughput, distributed data processing applications. Migrated workloads from Hive to Spark for performance improvement.
* Skilled in visual analytics and dashboarding using **Power BI, Tableau, Plotly**, and **Dataiku**, enabling business stakeholders with actionable insights and advanced data storytelling.
* Extensive experience with **NoSQL databases** like **Cassandra and HBase**, integrated into streaming and batch AI/ML pipelines.
* Well-versed in big data processing using **Hive, Pig, Impala, and Drill**, with the ability to write and optimize **custom UDFs** and complex queries.
* Strong understanding of **MapReduce**, **Hadoop architecture**, and components like **HDFS, YARN, NameNode, and ResourceManager** across **cloud-integrated** big data ecosystems.
* Hands-on experience with **GCP-native AI/ML tools** and **AutoML**, enabling quick prototyping and deployment of models in production environments.
* Experienced in Agile teams, collaborating with cross-functional stakeholders including Data Scientists, ML Engineers, DevOps, and Product teams to build **cloud-native AI/ML systems** that scale
* Good understanding of Data ingestion, Airflow Operators for Data Orchestration, and other related **python libraries**.
* Experience in designing Data Marts by following Star Schema and Snowflake Schema Methodology.
* Highly skilled in Business Intelligence tools like Tableau, PowerBI, Plotly and Dataiku.
* My expertise lies in selecting appropriate technologies, such as Apache Kafka, Apache Spark, or Apache Airflow, and architecting efficient data flow strategies to meet specific business requirements.
* Expertise in collecting, processing, and aggregating large amounts of streaming data using Kafka, Spark Streaming.
* Strong experience working in Hadoop eco-system integrated to the Cloud platform provided by **AWS** with several services like **Amazon EC2 instances, S3 bucket and RedShift**.
* Experienced **Azure Data Engineer with a strong specialization in AI/ML integration**, and also experience in designing and developing scalable big data pipelines and cloud-native analytics platforms.
* Experience in designing and developing applications in Spark using Python to compare the performance of Spark with Hive.
* Experienced 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 Databricks.
* Extensive experience designing and implementing cloud-based data aggregation systems with a strong focus on HIPAA compliance.
* Strong Knowledge on designing & building the data model for Snowflake cloud data warehouse.
* Strong Experience with AWS services like **Amazon EC2, S3, EMR, Amazon RDS, VPC**, Amazon Elastic Load Balancing, IAM, Auto Scaling, Cloud Front, CloudWatch, and Lambda to trigger resources
* Designed and implemented data pipelines using technologies such as Hadoop, Spark, and Kafka.
* Experience in Spark-Scala programming with good knowledge on Spark Architecture and its In-memory Processing.
* Hands on experience on Data Analytics Services such as Athena, Glue Data Catalog & Quick Sight
* Hands-on experience in **AI/ML**, including implementing **LLMs** for Retrieval-Augmented Generation (RAG), classification, and data feature extraction.
* Excellent understanding and knowledge of NOSQL databases like HBase, and Cassandra.
* Experienced in converting Hive/SQL queries into Spark transformations using Spark Data Frames and Python.
* Extensive experience working with spark distributed Framework involving Resilient Distributed Datasets (RDD) and Data Frames using Python, Scala and Java8.
* Involving in developing applications on **Windows, UNIX**, and **Linux Platforms**.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Ecosystems** | Hadoop, MapReduce, HDFS, HBase, Zookeeper, Hive, Pig, Sqoop, Cloudera, Hortonworks, Yarn, Cassandra, Oozie, Storm, and Flume. |
| **Streaming Technologies** | Spark, Kafka, Storm |
| **Scripting/Programming Languages** | Cassandra, Python, Scala, Regular Expressions, Shell scripting, PL/SQL, R, PySpark and Bash, Java, SQL, Java Scripting, HTML, CSS. |
| **Databases** | Data warehouse, RDBMS, NoSQL (Certified MongoDB), Oracle, HBase, Snowflake, MySQL. |
| **Java/J2EE Technologies** | Servlets, JSP (EL, JSTL, Custom Tags), JSF, Apache Struts, Junit, Hibernate 3.x, Log4J Java Beans, EJB 2.0/3.0, JDBC, RMI, JMS, JNDI. |
| **Tools** | Eclipse, JDeveloper, MS Visual Studio, Microsoft Azure HDInsight, Microsoft Hadoop cluster, JIRA, NetBeans, Eclipse. |
| **Methodologies** | Agile, Scrum, Waterfall |
| **Operating Systems** | Unix/Linux, Windows |
| **Machine Learning Skills (MLlib)** | Feature Extraction, Dimensionality Reduction, Model Evaluation, Clustering |
| **Cloud** | AWS, MS Azure, GCP |

**Professional Experience:**

**Client name: T-Mobile FEB 2023 to Present**

**Role: Senior AI/ML & Azure Data Engineer**

**Project description:** As part of T-Mobile's enterprise data modernization initiative, contributed to building and enhancing large-scale, cloud-native data engineering and AI solutions. Designed and implemented end-to-end data pipelines leveraging Azure Data Factory, Databricks, and Spark to support data ingestion, transformation, and analytics from multiple structured and unstructured sources. Spearheaded the integration of machine learning models into big data workflows using PySpark, MLlib, and TensorFlow for customer behavior prediction, operational forecasting, and intelligent automation.

**Responsibilities:**

* Developed in complete BigData flow of the application starting from data ingestion from upstream to HDFS.
* Developed and maintained scalable data pipelines using **Java**, **Spring Boot**, and **Kafka** to support real-time and batch processing needs.
* **Developed and maintained CI/CD pipelines** for PySpark-based ETL and data processing applications, ensuring seamless integration and deployment.
* Implemented RESTful APIs and integrated with data services to support seamless data flow across systems.
* Optimized Java code for performance, reducing memory usage and improving processing speed of large datasets.
* Working knowledge on Azure cloud components (Databricks, Data Lake, Blob Storage, Data Factory, Storage Explorer, SQL DB, SQL DWH, CosmosDB).
* Conducted quantitative analysis on cash products, ETD, and OTD to identify market trends, patterns, and opportunities for optimization.
* **Implemented testing and validation frameworks** within CI/CD pipelines to ensure data accuracy and workflow reliability.
* Integrated external data sources (SQL, Snowflake) into **Dataiku** projects to enable seamless data access and transformation.
* Integrated **machine learning models** into data pipelines using **Azure Databricks, Spark MLlib, and TensorFlow**, enabling large-scale **predictive analytics** for customer behavior and operational forecasting.
* Created and deployed **RESTful endpoints** for serving AI models in production using **Azure Kubernetes Service (AKS)** and **Docker containers**, enabling scalable and secure inference.
* Integrated **machine learning models** into data pipelines using **Azure Databricks, Spark MLlib, and TensorFlow**, enabling large-scale **predictive analytics** for customer behavior and operational forecasting
* Builted **data pipelines** to feed structured and unstructured data into **AI workloads**, incorporating **text analytics** and **NLP** techniques for parsing customer support logs and generating insights.
* Developed Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Sql Activity.
* Prepared dashboards using Tableau for summarizing Configuration, Quotes, Orders and other e - commerce data.
* Built and maintained access control mechanisms using IAM policies and role-based access to restrict PHI access.
* Extract, transform, and load (ETL) data from multiple federated data sources (JSON, relational database, etc.) with Data Frames in Spark.
* Developed XML-based payload generation and transformation logic for inter-service communication.
* Integrated legacy JMS/MQ-based messaging with AWS Lambda-based services to enable hybrid processing models.
* Coordinated with cross-functional teams to guide new AWS practitioners on best practices in serverless development and CI/CD.
* Worked with cloud-to-on-prem connectivity using VPC peering and secure tunnels for data exchange with legacy systems.
* Developed SparkAPI of queries against tables in entrise data warehouse in Azure Synapse Analytics by using table partitions.
* Worked on Schedule refresh in Power BI Service based on timings provided for all types of source datasets
* Integrated Python with big data technologies such as Apache Spark or Hadoop for scalable and distributed data processing.
* Implemented data ingestion pipelines using Snow Pipe for seamless and automated loading of data into Snowflake, enhancing overall data processing efficiency.
* Transformed and Copied data from the JSON files stored in a Data Lake Storage into an Azure Synapse Analytics table by using Azure Databricks
* Used Python to connect to various databases (e.g., SQL, NoSQL) for querying, updating, and managing data storage.
* Proficient in PySpark, Python, and SQL for data engineering and ML model development.
* 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, Devops write-back tool and backwards.
* Leveraged **Spark MLlib** for clustering and classification models applied to user activity logs and IoT telemetry, enhancing proactive alerting and recommendation systems.
* Deep understanding of cloud-based data solutions, distributed computing, and scalable ETL architectures in AWS environments.
* Developed end-to-end data pipelines in **Dataiku** for ETL processes, model training, and result visualization.
* Built and deployed data workflows using **Dataiku DSS**, integrating data preparation, machine learning, and reporting tasks in a unified environment.
* Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Collaborated with cross-functional teams to assess security risks and enforce best practices for PHI handling.
* Created data bricks notebooks using Python (PySpark), Scala and Spark SQL for transforming the data that is stored in Azure Data Lake stored Gen2 from Raw to Stage and curated zones.
* Deployed Apache Spark clusters on Kubernetes for distributed data processing.
* Integrated machine learning models into big data pipelines using Apache Spark MLlib and TensorFlow, enabling predictive analytics and pattern recognition on large-scale datasets.
* Led migration and modernization of healthcare data infrastructure while ensuring adherence to HIPAA and HITRUST security frameworks.
* Developed interactive data visualizations and dashboards using tools such as Apache Superset and Tableau, providing actionable insights and data-driven decision-making capabilities to stakeholders.
* Managed data lake storage and performed data cataloging using Delta Lake on Databricks, ensuring data consistency and governance.
* Orchestrated and deployed data processing applications, ETL (Extract, Transform, Load) jobs, or data pipelines in Kubernetes using containerization technologies like Docker.
* Developed robust data encryption strategies for data at rest and in transit to comply with HIPAA security regulations.
* Developed automated compliance auditing and reporting frameworks to support HIPAA regulatory requirements.
* Worked with Azure BLOB and Data lake storage and loading data into Azure SQL Synapse Analytics (DW).
* Worked on the creation of custom Docker container images, tagging and pushing the images and Docker consoles for maintaining the application of life cycle.
* Performed Data Visualization and Designed Dashboards with Tableau and generated complex reports including chars, summaries, and graphs to interpret the findings to the team and stakeholders.
* Experience in implementing Change Data Capture (CDC) using audit fields to track and update data efficiently.
* Automated report generation and distribution using Tableau Server's scheduling and subscription features, enhancing efficiency in data reporting.
* Strong Experience in Data Migration from RDBMS to Snowflake cloud data warehouse.

**Environment**: HDFS, Devops, NoSQL, Yarn, Mapreduce, Hive, Sqoop, Flume, Oozie, Quilk Replica, Azure Data Factory, HBase, Kafka, Impala, SparkSQL, Spark Streaming, Eclipse, Jira, Scala, JSON, Oracle, Teradata, CI/CD, PL/SQL UNIX Shell Scripting, Cloudera, MS Azure.

**Client Name: Centene corporation July 2020-Jan 2023**

**Role: Senior GCP Data Engineer**

**Project description:** At Centene, contributed to the modernization of enterprise healthcare analytics by designing and deploying scalable data engineering solutions on **Google Cloud Platform (GCP)**. Led the development of robust ETL/ELT pipelines using **Databricks, BigQuery, Dataflow**, and **Cloud Storage** to support ingestion, processing, and transformation of large-scale structured and semi-structured healthcare data. Migrated legacy Hadoop systems to GCP to improve scalability, performance, and reliability. Worked on the **end-to-end design of Snowflake-based data warehousing solutions**, including data modeling, transformation logic, and row-level security using authorized views. Collaborated with data science teams to integrate **Spark MLlib and PySpark** models into analytical workflows, enabling predictive insights on healthcare claims and operational trends. Implemented **CI/CD pipelines for dbt transformations**, improving deployment automation and version control of data artifacts.

**Responsibilities:**

* Developed 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.
* Involved in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinating tasks among the team.
* Designed and implemented end-to-end machine learning workflows for NLP use cases such as sentiment analysis, named entity recognition, and intent classification.
* Designed reusable data flows in Dataiku DSS to support scalable analytics and repeated business use
* Deployed NLP services using **FastAPI**, **Docker**, and **Kubernetes**, ensuring scalable and secure access to inference APIs.
* Integrated NLP components into production data pipelines using **Apache Spark**, **Airflow**, and **Kafka**, enabling real-time text processing and analytics.
* Implemented **Retrieval-Augmented Generation (RAG)** pipelines using **LLMs** (e.g., GPT, BERT) to improve response relevance in document-based question answering systems.
* Designed and optimized data warehouses on GCP using BigQuery, facilitating efficient analytics and reporting.
* Executed advanced SQL queries in GCP BigQuery, enhancing data processing capabilities.
* Used **Dataiku Plugins** and APIs to extend platform capabilities and connect with cloud-native services.
* Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target snowflake database.
* Processed Avro-based streaming data files as part of Kafka/Spark ingestion pipelines.
* Worked on SQL Server Integration Services (SSIS) to integrate and analyze data from multiple heterogeneous information sources.
* Implemented automated workflows using Dataiku Scenarios to schedule, monitor, and manage data tasks.
* Developed custom reports using SQL Server Reporting Services (SSRS) to review score cards, business trends based on the data from different Business or market locations.
* Designed and developed multiple metrics using 3 sigma control charts, P-Charts and Box plots in QlikView with dynamic subgrouping to analyze the shift in the sensor data
* Experience in working with healthcare datasets while ensuring compliance with regulatory requirements such as HIPAA.
* Designed and implemented data pipelines to ingest, process, and analyze healthcare claims data from multiple sources, ensuring data accuracy, integrity, and compliance with industry standards
* Leveraged PL/SQL for complex data analysis tasks, supporting the generation of insightful reports and analytics.
* Utilized Java frameworks for building scalable ETL (Extract, Transform, Load) pipelines, ensuring smooth data flow and transformation
* Extensive experience with NoSQL databases like HBase and Cassandra for scalable data storage solutions.
* Proficient in optimizing ETL processes for performance and scalability using ODI performance tuning
* Designed and developed Spring Boot microservices for efficient data processing and transformation.
* Got involved in migrating on prem Hadoop system to using GCP (Google Cloud Platform).
* Worked on a direct query using PowerBI to compare legacy data with the current data and generated reports and stored dashboards.
* Building multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate tasks among the team.
* Implemented Python scripts for real-time data processing, handling streaming data from sources like Kafka or Apache Flink.
* Worked on analyzing Hadoop clusters and different big data analytic tools including Pig, Hive.
* Used cloud shell SDK in GCP to configure the services Data Pros, Storage, BigQuery.
* Experience in learning architecting data intelligence solutions around Snowflake Data Warehouse and architecting snowflake solutions as developer.
* Deployed Dbt models and transformations to production environments using version control systems and continuous integration/continuous deployment (CI/CD) pipelines, enabling seamless and reliable deployment of data artifacts.
* Created GCP Big Query authorized views for row level security or exposing the data to other teams.
* Used ETL to implement the Slowly Changing Transformation, to maintain historical data in the Data warehouse.
* Performed ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Implemented data monitoring solutions within Spring Boot applications.
* Designed SSIS Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for the SSAS cubes (OLAP), SQL Server Reporting Services (SSRS).
* Processed and loaded bound and unbound data from Google Pub/Sub topics to BigQuery using Cloud Dataflow with Python.
* Created visual dashboards in **Tableau** and **Power BI** to present NLP outputs and trends to business teams.
* Implemented end-to-end data pipelines on Databricks to collect, process, and store large volumes of data from diverse sources, resulting in improved data accessibility and reliability.
* Implemented data integration solutions using Java to seamlessly connect diverse data sources and formats.
* Involved in migrating on-prem Hadoop systems to GCP (Google Cloud Platform).
* Designed and implemented data solutions on Google Cloud Platform, leveraging services such as BigQuery, Dataflow, and Storage to enable scalable and efficient data processing.
* Performed SQL Joins among Hive tables to get input for Spark batch process.
* Worked with a data science team to build statistical models with Spark MLLIB and PySpark.

**Environment:** Hadoop, HDFS, Hive, Map Reduce, Impala, Python, SSIS, PowerShell, SparkSQL, Kafka, Spark, Snowflake, Mapreduce, Power BI, PySpark, Spark MLlib, SQL, NoSQL, GCP.

**Client : Bank of America- Charlotte,NC November 2016 to June 2019**

**Senior Data Engineer**

**Project description:** Led the design and implementation of scalable, cloud-based data engineering solutions to support enterprise-level financial and clinical analytics initiatives. Developed robust data pipelines using Azure Data Factory, Databricks, Data Lake, and HDInsight, facilitating ingestion and transformation of large volumes of structured and unstructured data from diverse internal and external sources.Built and maintained data processing frameworks using Scala, Python, and Spark, enabling real-time analytics, fraud detection, and financial trend modeling. Created and deployed scalable ETL workflows in ADF, leveraging SQL activities and JSON-based pipeline automation to ensure rapid data integration across Azure components.

**Responsibilities:**

* Designed end to end scalable architecture to solve business problems using various Azure Components like HDInsight, Data Factory, Data Lake, Storage and Machine Learning Studio.
* specialize in designing and implementing clinical data warehouses that consolidate patient data from multiple sources for comprehensive analysis and reporting.
* Worked on Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, and Azure DW) and processing the data in In Azure Databricks.
* Used Azure Databricks for fast, easy, and collaborative spark-based platform on Azure.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Developed data processing algorithms using Scala and Python, driving actionable insights and decision-making in financial analytics.
* Automated system administration tasks and data workflows with Unix Shell Scripting, improving operational efficiency and reliability.
* Wrote multiple Hive UDFS using Core Java and OOP concepts and spark functions within Python programs.
* Undertook data analysis and collaborated with down-stream, analytics team to shape the data according to their requirement.
* Managed Snowflake and MongoDB databases, ensuring data integrity, security, and scalability.
* Proficient in SQL for querying and managing relational databases, facilitating data analysis and reporting.
* Experienced in HiveQL for big data processing, enabling effective management of large datasets.
* Utilized Scala for scalable data processing solutions, improving overall system performance and reliability.
* Used Azure Event Grid for managing event service that enables you to easily manage events across many different Azure services and applications.
* Delta lake supports merge, update and delete operations to enable complex use cases.
* Used Azure Data Catalog which helps in organizing and to get more value from their existing investments.
* Used Azure Synapse to bring these worlds together with a unified experience to ingest, explore, prepare, manage, and serve data for immediate BI and machine learning needs.

**Environment:** Azure (Storage, DW, ADF, ADLS, Databricks), AWS Redshift, Ubuntu 16.04, Hadoop 2.0, Spark (PySpark, Spark streaming, SparkSQL, SparkMLlib), Mapreduce, Hive, Nifi, Jenkins, Pig 0.15, Python 3.x (Nltk, Pandas), AI/ML, Tableau 10.3, Git, Talend.

**Client: Genpact – India August 2014 - March 2016**

**Big Data Engineer**

**Project description**: Contributed to the development of large-scale big data solutions aimed at social media analytics and enterprise reporting by leveraging the Hadoop ecosystem and AWS cloud services. Designed and implemented end-to-end data pipelines for ingesting, processing, and analyzing high-volume datasets from multiple sources including Facebook, Twitter, SQL Server, MySQL, and Teradata, using tools such as Flume, Sqoop, and Shell scripting.Built robust ETL workflows using Hive, Spark, and Oozie, applying best practices like dynamic partitioning, bucketing, and compression to enhance performance and storage efficiency. Migrated legacy HiveQL scripts to Spark SQL, significantly improving query execution times and enabling real-time analytics over large datasets. Utilized Zeppelin notebooks and AWS Athena for interactive querying and reporting, and integrated data into AWS Redshift and HBase for scalable storage and analytics.

**Responsibilities:**

* Extracted feeds form social media sites such as **Facebook, Twitter using Python scripts**.
* Developed in running all the hive scripts through hive. Hive on Spark and some through **Spark SQL**.
* Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
* Developed in complete Bigdata flow of the application starting from data ingestion from upstream to HDFS, processing and analyzing the data in HDFS.
* Implemented reporting in PySpark, Zeppelin & querying through Airpal & AWS Athena.
* Proven experience with **ETL frameworks** (Airflow, Luigi, or our own open sourced garcon)
* Created Hive schemas using performance techniques like partitioning and bucketing.
* Created data models for **AWS Redshift**and **Hive** from dimensional data models.
* Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
* Developed Star and **Snowflake** schemas based dimensional model to develop the data warehouse
* Build machine learning models to showcase Big data capabilities using **PySpark** and **MLlib**.
* Worked on Amazon AWS concepts like **EMR**and**EC2** web services for fast and efficient processing of Big Data.
* Developed in loading data into HBase NoSQL database.
* Building, Managing and scheduling Oozie workflows for end-to-end job processing.
* Worked on Hortonworks-HDP 2.5distribution.
* Responsible for building-scalable distribution data solution using Hadoop.
* Built **PL/SQL** (Procedures, Functions, Triggers, and Packages) to summarize the data to populate summary tables that will be used for generating reports with performance improvement.
* Involved in importing data from **MS SQL Server,** **MySQL** and **Teradata** into **HDFS** using **Sqoop.**
* Played a key role in dynamic partitioning and Bucketing of the data stored in **Hive Metadata**.
* Wrote **Hive** QL queries for integrating different tables for create views to produce result set.
* Collected the log data from **Web Servers** and integrated into **HDFS** using **Flume**.
* Worked on loading and transforming of large sets of **structured** **and unstructured data**.
* Used **Map** Reduce programs for data cleaning and transformations and load the output into the **Hive tables** in different file formats.
* Worked on extending Hive and Pig core functionality by writing custom UDFs using **Java**.
* Analyzing of Large volumes of structured data using **Spark SQL**.
* Migrated **Hive QL** queries into **Spark SQL** to improve performance.

**Environment:** Hortonworks, Hadoop, HDFS, Pig, Sqoop, Hive, Oozie, Zookeeper, NoSQL, HBase, Shell Scripting, Scala, Spark, Spark SQL.