**Sr. Data Engineer**

**Name: Mounika**

**Ph no: 8602811711**

**Email: monikayesh2023@gmail.com**

**Professional Summary:**

* Around **10 years** of professional experience as a Software developer in design, development, deploying and supporting large scale distributed systems.
* Extensive experience in Amazon Web Services (AWS) Cloud services such as EC2, VPC, S3, IAM, EBS, RDS, ELB, VPC, Route53, Ops Works, DynamoDB, Autoscaling, CloudFront, CloudTrail, CloudWatch, CloudFormation, Elastic Beanstalk, AWS SNS, AWS SQS, AWS SES, AWS SWF & AWS Direct Connect.
* Extensive experience in providing solutions for Big Data using Hadoop, Spark, HDFS, Map Reduce, YARN, Kafka, Pig, Hive, Sqoop, HBase, Oozie, Zookeeper, Cloudera Manager, Horton works.
* Strong experience working with Amazon cloud services like EMR, Redshift, **DynamoDB**, **Lambda**, Athena, Glue, S3, API Gateway, RDS, CloudWatch for efficient processing of Big Data.
* Experience in developing web applications by using Python, Django, C++, XML, CSS, HTML, JavaScript and jQuery.
* Experience in analyzing data using Python, R, SQL, Microsoft Excel,Hive, PySpark, Spark SQL for Data Mining, Data Cleansing, Data Munging and Machine Learning.
* Hands-on experience building **PySpark**, Spark Java and Scala applications for batch and stream processing involving Transformations, Actions, Spark SQL queries on RDD’s, Data frames.
* Strong experience writing, troubleshooting, and optimizing Spark scripts using Python, Scala.
* Experienced in using Kafka as a distributed publisher-subscriber messaging system.
* Strong knowledge on performance tuning of Hive queries and troubleshooting various issues related to Joins, memory exceptions in Hive.
* Designed, built, and deployed a multitude application utilizing almost all AWS stack (Including EC2, R53, S3, RDS, HSM Dynamo DB, SQS, IAM, and EMR), focusing on high-availability, fault tolerance, and auto-scaling
* Implemented AWS X-Ray allows you to visually detect node and edge latency distribution directly from the service map Tools, like Splunk, Sumologic can be used for log analysis but when comes to distributed tracing with in the AWS, X-Ray will be provided much better features with service map, traces with in depth analysis with minimal configuration with not much maintenance.
* Hands-on experience in handling database issues and connections with SQL and NoSQL databases like MongoDB, Cassandra, Redis, CouchDB, DynamoDB by installing and configuring various packages in python.
* Exceptionally good understanding of partitioning, bucketing concepts in Hive and designed both Managed and External tables in Hive.
* Experience in importing and exporting data between HDFS and Relational Databases using Sqoop.
* Experience working with RDBMS including Oracle/ DB2, SQL Server, PostgreSQL 9.x, MS Access and Teradata for faster access to data on HDFS.
* Extensive Design, implement and test services experience in aMicro Service-oriented architecture and has implemented the REST based Microservices using theSpringBoot,SpringData with JPA (Java Persistence API).
* Experience in real time analytics with Spark Streaming, Kafka and implementation of batch processing using Hadoop, Map Reduce, Pig and Hive.
* Experienced in building highly scalable Big-data solutions using NoSQL column-oriented databases like Cassandra, MongoDB and HBase by integrating them with Hadoop Cluster.
* Manager and SaaS, PaaS and IaaS concepts of Cloud Computing and Implementation Worked with
* Extensive work on ETL consisting of data transformation, data sourcing, mapping, conversion and loading data from heterogeneous systems like flat files, Excel, Oracle, Teradata, MSSQL Server.
* Experienced in Vue.JS, TypeScript or Node.JS PostgreSQL and building Data Pipeline in AWS including ETL (Spark, Glue, etc.)
* Experience of building ETL production pipelines using Informatica Power Center, SSIS, SSAS, SSRS.
* Proficient at writing MapReduce jobs and UDF’s to gather, analyse, transform, and deliver the data as per business requirements and optimize the existing algorithms for best results.
* Experience in working with Data warehousing concepts like **Star** **Schema**, **Snowflake Schema**, DataMarts, Kimball Methodology used in Relational and Multidimensional data modelling.
* Strong experience leveraging different file formats like Avro, ORC, Parquet, JSON and Flat files.
* Sound knowledge on Normalization and Denormalization techniques on **OLAP** and **OLTP** systems.
* Practical experience in with PostgreSQL (MVCC, WAL, OOP, Foreign Data Wrapper) and MySQL & MariaDB
* Good experience with Version Control tools Bitbucket, GitHub, GIT.
* Experience with Jira, Confluence, and Rally for project management and Oozie, AirFlow scheduling tools.
* Monitor portfolio performance, attribution, and benchmarks, making adjustments as needed to achieve targeted returns.
* Conduct thorough research and analysis of financial markets, macroeconomic trends, and industry sectors to provide actionable insights for trading decisions.
* Monitor news, economic indicators, and geopolitical events to identify potential market opportunities and risks.
* Good experience with **Quantexa** platform updates, **AWS** services, and best practices to enhance data engineering capabilities and contribute to continuous improvement initiatives.
* Integrate data and workflows into the **Quantexa platform**, ensuring seamless connectivity and data synchronization for advanced analytics and entity resolution.
* Configure and manage connectors, APIs, and data loaders within the **Quantexa** environment
* Experienced in Strong scripting skills in Python, Scala and UNIX shell.
* Involved in writing Python, Java API’s for Amazon Lambda functions to manage the **AWS** services.
* Experience in design, development and testing of Distributed Client/Server and Database applications using **Java**, **Spring**, **Hibernate**, JSP, JDBC, **REST** services on Apache Tomcat Servers.
* Hands on working experience with **RESTful** **API’s**, API life cycle management and consuming RESTful services.
* Have good working experience in Agile/Scrum methodologies, communication with scrum calls for project analysis and development aspects.
* Worked with Google Cloud (GCP) Services like Compute Engine, Cloud Functions, Cloud DNS, Cloud Storage and Cloud Deployment Manager and **SaaS**, **PaaS** and **IaaS** concepts of Cloud Computing and Implementation using GCP.

**Technical Skills:**

**Programming Languages:** Python, Scala, SQL, Java, C/C++, Shell Scripting

**Web Technologies:** HTML, CSS, XML, AJAX, JSP, Servlets, JavaScript

**Big Data Stack:** Hadoop, Spark, MapReduce, Hive, Pig, Yarn, Sqoop, Flume, Oozie, Kafka, Impala, Storm

**Cloud Platform:** Amazon Web Services (**AWS**), Google Cloud Platform (GCP),

**Relational databases**: Oracle, MySQL, SQL Server, DB2, PostgreSQL, Teradata, Snowflake, Splunk

**NoSQL databases:**  MongoDB, Cassandra, HBase, Pig

**Version Control Systems**: Bitbucket, GIT, SVN, GitHub

**IDEs**: PyCharm, IntelliJ IDEA, Jupyter Notebooks, Google Colab, Eclipse

**Operating Systems**: Unix, Linux, Windows

**Professional Experience:**

**Client: Pet Smart, CT Nov 2023 – Till Date**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Worked on building the data pipelines **(ELT/ETL Scripts)**, extracting the data from different sources

(DB2, **AWS** S3 files), transforming and loading the data to **the Data Warehouse** **(AWS** **Redshift).**

* Worked on adding the **Rest API l**ayer to the **ML models built using Python, Flask & deploying the models in AWS BeanStalk Environment using Docker containers.**
* Ability to debug and troubleshoot **Terraform** deployments.
* Worked on developing & adding few Analytical dashboards using Looker product.
* Designed and implemented scalable and reliable data pipelines using Kubernetes and Apache Spark for real-time data processing
* Developed rest API's using python with flask and Django framework and done the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Utilized **Kubernetes API** and Helm charts to automate deployment and monitoring of data applications
* Written Terraform scripts to automate AWS services which include ELB, CloudFront distribution, RDS, EC2, database security groups, Route 53, VPC, Subnets, Security Groups, and S3 Bucket and converted existing AWS infrastructure to AWS Lambda deployed via Terraform and AWS CloudFormation.
* Automate Datadog Dashboards with the stack through Terraform Scripts.
* 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.
* Implemented data security measures in Kubernetes clusters to ensure data privacy and compliance with regulations
* Developed an **AWS TDM (Test Data Management)** strategy for the organization, reducing the time and effort required for data provisioning and ensuring data privacy and security.
* Implemented AWS data quality checks using **AWS DQ (Data Quality)** tools, detecting and resolving data issues in real-time, resulting in improved data accuracy and consistency.
* Used AWS data pipeline for Data Extraction, Transformation and Loading from homogeneous or heterogeneous data sources and built various graphs for business decision-making using Python matplot library
* Worked on building the aggregate tables & de-normalized tables, populating the data using **ETL** to improve the looker analytical dashboard performance and to help data scientist and analysts to speed up the **ML** model training & analysis.
* Created New **Dashboard**s, reports, scheduled searches and alerts using **spunk**.
* Integrate Pager duty with **Splunk** to generate the Incidents from **Splunk**.
* Developed custom **Jenkins’s jobs/pipelines** that contained Bash **shell scripts** utilizing the **AWS** **CLI** to automate infrastructure provisioning.
* Good Understanding of other AWS services like S3, EC2 IAM, RDS Experience with Orchestration and Data Pipeline like AWS Step functions/Data Pipeline/Glue
* Developed a user-eligibility library using **Python** to accommodate the partner filters and exclude these users from receiving the credit products.
* Conducted performance testing and troubleshooting of Kubernetes clusters to identify and resolve bottlenecks
* Extensively involved in Cluster planning, Installation, troubleshooting and Performance Tuning of the Mongo and PostgreSQL instances on premise and on cloud.
* Used Maven to manage project dependencies and build automated deployment processes to deploy data pipelines and applications into production environments.
* Developed and maintained data warehouse using Snowflake to enable ad-hoc querying and reporting for business users
* Good Understanding of other **AWS services** like **S3, EC2 IAM, RDS** Experience with Orchestration and Data Pipeline like **AWS Step functions/Data Pipeline/Glue**.
* Hands-on experience on working with AWS services like Lambda function, Athena, DynamoDB, **Step functions, SNS, SQS, S3, IAM etc.**
* Integrated lambda with **SQS and DynamoDB** with **step functions** to iterate through list of messages and updated the status into **DynamoDB table**.
* Utilized JBoss Data Virtualization (JDV) to create a virtualized data layer, providing a unified view of disparate data sources to support
* Automated data ingestion and processing using JBoss Drools and Apache Airflow, reducing manual effort and increasing efficiency
* Hands-on experience on working with AWS services like Lambda function, Athena, DynamoDB, Step functions, SNS, SQS, S3, IAM etc.
* Creation of indexes, forwarder & indexer management, Splunk Field Extractor IFX, Search head Clustering, Indexer clustering, Splunk upgradation.
* Install and configured Splunk clustered search head and Indexer, Deployment servers, Deployers.
* Designing and implementing Splunk - based best practice solutions.
* Design, setup and Administer high volume and High transaction (failover + offsite) PostgreSQL DB installations with streaming replication.
* Utilized WebLogic JDBC data sources to connect and integrate external databases, enabling seamless data integration and business continuity.
* Implemented security best practices for **WebLogic** environments, including SSL configuration, dynamic user authentication, and authorization, to ensure data confidentiality and integrity
* Collaborated with development teams to troubleshoot and resolve production issues related to WebLogic, ensuring minimal downtime and maximum availability
* Designed and implemented a highly scalable and automated data pipeline using Ansible for data ingestion, transformation, and loading.
* Developed and maintained **Ansible playbooks** to deploy and manage big data clusters on cloud platforms like **AWS, Google Cloud, and Azure**.
* Implemented complex data integration processes using **Ansible** to connect and transfer data between different systems and databases.
* Utilized **Ansible's dynamic** inventory feature to dynamically manage and provision resources in the data pipeline based on workload demands.
* Automated deployments of data warehouse applications and BI tools using **Ansible**, reducing deployment time by 50%.
* Compiling AHC reports and implementing changes on Postgresql/PPAS clusters.
* Handling large databases in production support environment for various CMM 5 level customers for disaster.
* Conditions.
* Maintenance and management of in-house PostgreSQL databases.
* Designed disaster recovery processes for data applications running on Kubernetes to ensure minimal downtime
* Creating and managing user management tasks on Mongodb and PostgreSQL.
* Implemented logging and monitoring solutions, such as Prometheus and ELK stack, for Kubernetes clusters to track performance and detect issues
* Built the data pipelines to aggregate the user click stream session **data using** **spark** streaming module which reads the clickstream data from Kinesis streams and store the aggregate results in **S3** and data and eventually loaded to **AWS** **Redshift** warehouse.
* Worked on supporting & building the infrastructure for the core module of the Credit Sesame i.e Approval Odds, started with Batch ETL, moved to micro-batches and then converted to a real time predictions.
* Implementation of end-to-end data solution on **Azure using Azure Databricks**, **ADF, DW and PowerBI**.
* Designed a robust data modelling environment using **Databricks on Azure**, enabling consumers to easily operate highly descriptive Notebooks in a fully governed environment
* Migration large data sets to **Databricks (Spark)**, create and administer cluster, load data, configure data pipelines, loading data from **ADLS Gen2 to Databricks using ADF pipelines**
* Extensive hands-on experience of writing notebooks in data bricks using python/Spark SQL for complex data aggregations, transformations, schema operations. Good familiarity with **Databrick**s delta and data frames concepts
* Developed the **AWS** **Lambda** server less scripts to handle ad-hoc requests.
* **Performed Cost optimization reduced the infrastructure costs**.
* Configure and manage connectors, APIs, and **data loaders** within the **Quantexa environment**
* Worked on building the data pipelines using **PySpark** (**AWS** **EMR**), processing the data files present in **S3** and loading it to **Redshift**.
* Other activities include supporting and keeping the **data pipelines active, working with Product Managers, Analysts, Data Scientist & addressing the requests coming from them, unit testing, load testing** and **SQL** optimizations in **DB2 server**.

**Environment**: Groovy, Python, Flask, Terraform, Numpy, Pandas, DB2, Splunk, Cassandra, **AWS** EMR, Spark, **AWS** Kinesis, **AWS** Redshift, **AWS** EC2, **AWS** S3, IAM, **AWS** BeanStalk, **AWS** Lambda, **AWS** data pipeline, Quantexa, **AWS** cloud-watch, Docker, Shell scripts, Looker.

**Client: Qualcomm, India Aug2021 – Mar 2023 Role: Big Data Developer**

**Responsibilities:**

* As a **Data Engineer** I am responsible for building scalable distributed data solutions using **Hadoop**.
* Involved in Agile Development process (Scrum and Sprint planning).
* Handled Hadoop cluster installations in Windows environment.
* Worked with Terraform Templates to automate the Azure Iaas virtual machines using terraform modules and deployed virtual machine scale sets in production environment**.**
* **Migrated on-premises** environment in **GCP (Google Cloud Platform)**
* Experience building and deploying cloud infrastructure using Terraform
* Worked in Server infrastructure development on AWS Cloud, extensive usage of Virtual Private Cloud (VPC), Cloud Formation, Lambda, Cloud Front, Cloud Watch, IAM, EBS, Security Group, Auto Scaling, Dynamo DB, Route53, and Cloud Trail.
* Used MongoDB to stored data in JSON format and developed and tested many features of dashboard using Python, Bootstrap, CSS, and JavaScript.
* Involved in Agile Development process (Scrum and Sprint planning).
* Involved in various sectors of business, with In-depth knowledge of SDLC (System Development Life Cycle) with all phases of Agile - Scrum, & Waterfall.
* Migrated data warehouses to Snowflake Data warehouse.
* Defined virtual warehouse sizing for **Snowflake** for different type of **workloads**.
* Demonstrated knowledge of **AWS**, **Azure, Google Cloud Platform, and other cloud providers**
* **Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) BigQuery.**
* Ability to design, develop, and implement **Terraform scripts** for **infrastructure automation**.
* **Proven understanding of the principles of Infrastructure as Code (IaC).**
* Worked with **Google Cloud (GCP**) Services like **Compute Engine, Cloud Functions, Cloud DNS, Cloud Storage and Cloud Deployment Manager and SaaS, PaaS and IaaS concepts of Cloud Computing and Implementation using GCP.**
* Involved in migration of an **Oracle SQL ETL** to run on **Google cloud platform** using **cloud Dataproc** & **BigQuery, cloud pub/sub for triggering the Apache Airflow jobs**.
* **Extracted data from data lakes, EDW to relational databases for analysing and getting more meaningful insights using SQL Queries in DB2 DB and PySpark.**
* Developed **PySpark** **script** to merge static and **dynamic files and cleanse the data**.
* Created **Pyspar**k procedures, functions, packages to load data.
* Designed, developed and did maintenance of **data integration** programs in a **Hadoop and RDBMS** environment with both traditional and non-traditional source systems.
* Knowledge and experience on using **Python Numpy**, **Pandas, Sci-kit Learn, Onnx & Machine Learning**.
* Developed **MapReduce** programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
* Use the RUP and agile methodology to conduct new development and maintaining software
* Check the data and tables structure in the PostgreSQL & Redshift databases and run the queries to generate report
* Monitored containers in AWS EC2 machines using Datadog API and ingest, enrich data into the internal cache system.
* Configured servers to send the server and application data to Splunk.
* Hands on experience with generating reports using Splunk.
* Manage and maintain optimize SQL Server, PostgreSQL in Windows, and Linux environment
* Good experience with Quantexa platform updates, AWS services, and best practices to enhance data engineering capabilities and contribute to continuous improvement initiatives.
* Integrate data and workflows into the **Quantexa** platform, ensuring seamless connectivity and data synchronization for advanced analytics and entity resolution.
* Wrote **Sqoop Scripts** for importing and exporting data from **RDBMS (DB2) to HDFS**.
* Set up **Data Lake in Google cloud using Google cloud storage, BigQuery and Big Table**.
* Developed scripts **in BigQuery** and connecting it to reporting tools.
* **Designed workflows using Airflow to automate the services developed for Change data capture.**
* Carried out data transformation and cleansing using **SQL** queries and **PySpark**.
* Used **Kafka** and **Spark streaming** to ingest real time or near real time **data** in **HDFS**.
* Worked related to downloading **BigQuery data** into **Spark data** frames for **advanced ETL capabilities**.
* Worked on **PySpark APIs** for **data transformations**.
* Built reports for monitoring **data loads into GCP and drive reliability at the site level**.
* Participated in daily stand-ups, bi-weekly scrums and PI panning.

**Environment:** Hadoop 3.3, GCP, BigQuery, PostgreSQL, Big Table, Spark 3.0, PySpark, Sqoop 1.4.7, ETL, HDFS, Snowflake DW, DB2, MapReduce, Kafka 2.8 and Agile process.

**Client: CITI , India** **Oct 2020 – June 2021**

**Role: Data Engineer**

**Responsibilities**

* Extensive experience in working with **AWS** **cloud Platform** (**EC2, S3, EMR, Redshift, Lambda and Glue**).
* Working knowledge of **Spark RDD**, **Dataframe API, Data set API, Data Source API, Spark SQL and Spark Streaming.**
* Developed **Spark Applications** by using **Python** and Implemented **Apache Spark data processing Project to handle data from various RDBMS and Streaming sources.**
* Worked with **Spark** for improving performance and optimization of the existing algorithms in **Hadoop**.
* Using **SparkContext**, **Spark-SQL, Spark MLlib, Data Frame, Pair RDD and Spark YARN**.
* Used **Spark Streaming APIs** to perform transformations and actions on the fly for building common.
* Learner **data model** which gets the data from **Kafka** in real time and persist it to **Cassandra**.
* Developed **Kafka** consumer **API in python for consuming data from Kafka topics**.
* **Consumed Extensible Markup Language (XML) messages using Kafka and processed the XML file using Spark Streaming to capture User Interface (UI) updates.**
* Developed Preprocessing job using **Spark Data** **frames** to flatten JSON documents to flat file.
* Load D-Stream data into **Spark RDD** and do in memory data Computation to generate output response.
* Involved in Agile development methodology active member in scrum meetings.
* Experienced in writing live Real-time Processing and core jobs using **Spark Streaming with Kafka as a Data pipeline system.**
* Migrated an existing on-premises application to **AWS**. Used **AWS** services like **EC2** and **S3** for data sets processing and storage.
* Experienced in Maintaining the Hadoop cluster on **AWS** **EMR**.
* Loaded data into **S3** buckets using **AWS Glue** and **PySpark**. Involved in filtering data stored in **S3** buckets using **Elasticsearch** and **loaded data into Hive external tables**.
* Configured Snow pipe to pull the data from **S3 buckets** into **Snowflakes table**.
* Stored incoming data in the **Snowflakes** staging area.
* Created numerous ODI interfaces and loaded into **Snowflake DB**.
* Worked on **Amazon Redshift** for shifting all **Data warehouses** into one **Data warehouse**.
* Good understanding of **Cassandra architecture**, replication strategy, gossip, snitches etc.
* **Designed columnar families in Cassandra and Ingested data from RDBMS, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.**
* Used the **Spark Data Cassandra Connector to load data to and from Cassandra**.
* Worked from Scratch in Configurations of **Kafka such as Mangers and Brokers**.
* **Experienced in creating data-models for Clients transactional logs, analyzed the data from Cassandra.**
* **Tables for quick searching, sorting and grouping using the Cassandra Query Language.**
* **Tested the cluster performance using Cassandra-stress tool to measure and improve the Read/Writes.**
* **Used Hive QL to analyse the partitioned and bucketed data, Executed Hive queries on Parquet tables.**
* Used **Apache Kafka** to aggregate web log data from multiple servers and make them available in downstream systems for Data analysis and engineering type of roles.
* Worked in **Implementing Kafka Security** and **boosting its performance.**
* **Experience in using Avro, Parquet, RCfile and JSON file formats, developed UDF in Hive.**
* Developed Custom **UDF** in **Python** and used **UDFs** for sorting and preparing the data.
* **Worked on Custom Loaders and Storage Classes in PIG to work on several data formats like JSON, XML, CSV and generated Bags for processing using pig etc.**
* Developed Oozie coordinators to **schedule Hive scripts** to **create Data pipelines**.
* **Written several Map Reduce Jobs using Pyspark, Numpy and used Jenkins for Continuous integration.**
* On cluster and testing of **HDFS, Hive, Pig and MapReduce to access cluster** for new users.
* **Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.**

**Environment:** Spark, Spark-Streaming, Spark SQL, **AWS** EMR, map R, HDFS, Hive, Pig, Apache Kafka, Sqoop, Python, Pyspark, Shell scripting, Linux, MySQL Oracle Enterprise DB, SOLR, Jenkins, Eclipse, Oracle, Git, Oozie, Tableau, MySQL, Soap, Cassandra and Agile Methodologies.

**Client: Dell, India Mar 2015 – Sept 2020**

**Role: Software Developer**

**Responsibilities:**

* Participate in requirement grooming meetings which involves understanding functional requirements from business perspective and providing estimates to convert those requirements into software solutions (**Design and Develop & Deliver the Code to IT/UAT/PROD and validate and manage data Pipelines from multiple applications with fast-paced Agile Development methodology using Sprints with JIRA Management Tool).**
* Responsible to check data in **DynamoDB** tables and to check **EC2** instances are upon running for
* (**DEV, QA, CERT and PROD)** in **AWS**.
* Analysis on existing data flows and create high level/low level technical design documents for business stakeholders that confirm technical design aligns with business requirements.
* Creation and deployment of Spark jobs in different environments and loading data to no sql database **Cassandra/Hive/HDFS**. Secure the data by implementing encryption based.
* Implemented **AWS** **solutions using E2C, S3, RDS, EBS, Elastic Load Balancer, Auto scaling groups, Optimized volumes, and EC2 instances and created monitors, alarms, and notifications for EC2 hosts using Cloud Watch.**
* **Developing code using Apache Spark and Scala, IntelliJ, NoSQL databases (Cassandra), Jenkins, Docker pipelines, GITHUB, Kubernetes, HDFS file System, Hive, Kafka for streaming Real time streaming data, Kibana for monitor logs etc. authentication/authorization to the data Responsible to deployments to DEV, QA, PRE-PROD (CERT) and PROD using AWS.**
* **Scheduled Informatica Jobs through Autosys scheduling tool.**
* **Created quick Filters Customized Calculations on SOQL for SFDC queries, Used Data loader for ad hoc data loads for Salesforce**
* Extensively worked on Informatica power centre Mappings, Mapping Parameters, Workflows, Variables and Session Parameters.
* Responsible for facilitating load data pipelines and benchmarking the developed product with the set performance standards.
* Used Debugger within the Mapping Designer to test the data flow between source and target and to troubleshoot the invalid mappings.
* **Worked on SQL tools like TOAD and SQL Developer to run SQL Queries and validate the data.**
* Study the existing system and conduct reviews to provide a unified review on jobs.
* Involved in Onsite & Offshore coordination to ensure the deliverables.
* Involving in testing the database using complex SQL scripts and handling the performance issues effectively.

**Environment:** Apache spark 2.4.5, Scala2.1.1, Cassandra, HDFS, Hive, GitHub, Jenkins, kafka, SQL Server 2008, Salesforce Cloud, Visio, TOAD, Putty, Autosys Scheduler, UNIX, **AWS**, WinScp, Salesforce data loader, SFDC Developer console.

**Client: Pfizer, India Jul 2014 – Fab 2016**

**Role: Software Developer**

**Responsibilities:**

* **Imported the data from various formats like JSON, Sequential, Text, CSV, AVRO and Parquet to HDFS cluster with compressed for optimization.**
* Worked on ingesting data from RDBMS sources like - **Oracle, SQL Server and Teradata into HDFS using Sqoop.**
* Loaded all datasets into Hive from Source **CSV** files using Spark and Cassandra from Source CSV files using Spark.
* **Created environment to access Loaded Data via Spark SQL, through JDBC ODBC (via Spark Thrift Server).**
* **Developed real time data ingestion/ analysis using Kafka / Spark-streaming.**
* **Configured Hive and written Hive UDF's and UDAF's Also, created Static and Dynamic with bucketing as required.**
* Worked on writing Scala programs using Spark on Yarn for analysing data.
* Managing and scheduling Jobs on a Hadoop cluster using Oozie.
* Created Hive External tables and loaded the data into tables and query data using HQL.
* Written Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Developed Oozie workflow for scheduling and orchestrating the ETL process and worked on Oozie workflow engine for job scheduling.
* Managed and reviewed the Hadoop log files using Shell scripts.
* Migrated ETL jobs to Pig scripts to do transformations, even joins and some pre-aggregations before storing the data onto HDFS.
* Using Hive join queries to join multiple tables of a source system and load them to Elastic search tables.
* Real time streaming, performing transformations on the data using Kafka and Kafka Streams.
* Built NiFidataflow to consume data from Kafka, make transformations on data, place in HDFS & exposed port to run Spark streaming job.
* Developed Spark Streaming Jobs in Scala to consume data from Kafka topics, made transformations on data and inserted to HBase.
* Implemented Spark using Scala and SparkSQL for faster testing and processing of data.
* Experience in managing and reviewing huge Hadoop log files.
* Collected the logs data from web servers and integrated in to HDFS using Flume.
* Expertise in designing and creating various analytical reports and Automated Dashboards to help users to identify critical KPIs and facilitate strategic planning in the organization.
* Involved in Cluster maintenance, Cluster Monitoring and Troubleshooting.
* Worked with Avro Data Serialization system to work with JSON data formats.
* Used Amazon Web Services (**AWS**) S3 to store large amounts of data in identical/similar repositories.

**Environment:** Spark, Spark SQL, Spark Streaming, Scala, Kafka, Hadoop, HDFS, Hive, Oozie, Pig, Nifi, Sqoop, **AWS** (EC2, S3, EMR), Shell Scripting, HBase, Jenkins, Tableau, Oracle, MySQL, Teradata and **AWS**.