**KOTESWAR K**

**SR GCP Data Engineer**

**Email ID:diana@fluxteksol.com**

**Contact: +8326122540**

**Professional Summary:**

* Over 9+ years of experience in **Analyzing, Designing, Developing, and Implementing data, architecture, and frameworks as a Data Engineer and Cloud technologies (AWS, GCP, AZURE**).
* Specialized in Data Warehousing and DecisionSupport Systems and extensive experience in implementing Full Life cycle **Data Warehousing Projects and in Hadoop/big data-related technology experience in Storage, Querying, Processing, and analysis of data.**
* Excellent knowledge on**Hadoop Architecture and ecosystems such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, and Map Reduce programming paradigm**.
* Knowledge in installing, configuring, and using **Hadoop ecosystem components like Hadoop Map Reduce, HDFS, HBase, Oozie, Hive, Sqoop, Zookeeper, and Flume.**
* Experience in analyzing data **using HiveQL, HBase and custom Map Reduce programs**.
* Experience in importing and exporting data using **Sqoop from HDFS to Relational Database Systems like Teradata, Oracle, SQL Server, and vice-versa.**
* An accomplished Data Engineer experienced in **ingestion, storage, querying, processing**, and analysis of big data, an expert in coming up with data warehousing solutions working with a variety of **database technologies**.
* Design & implement migration strategies for traditional systems on Azure (Lift and shift/Azure Migrate, other third-party tools) worked on Azure suite like **Azure SQL Database, Azure Data Lake (ADLS), Azure Data Factory (ADF) V2, Azure SQL Data Warehouse, Azure Service Bus, Azure Key Vault, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, AZURE data Platform Service.**
* Extensive Hands-on experience in **GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Data Proc, GKE, PUB/SUB, Apache beam, Cloud Composer, and Cloud SQL**.
* Build a program with **Python and Apache beam and execute it in cloud Dataflow** to run Data validation between raw source file and big query tables.
* Developed Apache **Spark jobs using Python in a test environment** for faster data processing and used **Spark SQL for querying**.
* Experienced in **Spark Core, Spark RDD, Spark Streaming, Data Frames, Pair RDD, Spark MLIB, and Spark Deployment** Architectures.
* Implemented POC to migrate map-reduce jobs **into Spark transformations using Python**.
* Extensively worked on **Spark using Scala**, and Python on the cluster for computational analytics, installed it on top of Hadoop performed advanced analytical applications by making use of Spark with Hive and SQL/Oracle.
* Build a program with Python and **Apache beam** and execute it in cloud Dataflow to run Data validation between raw source file and **Big query table**s.
* Experienced with **Docker and Kubernetes on multiple cloud providers**, from helping developers build and containerize their application (CI/CD) to deploying either on public or private cloud.
* Worked **with HBase to conduct quick look ups (updates, inserts and deletes) in Hadoop**.
* Experience **in Oozie and workflow** scheduler to manage Hadoop jobs **by Direct Acyclic Graph (DAG)** of actions with control flows
* Good understanding of Statistics and developing **Machine learning models such as Linear Regression AND Logistic Regression**
* Strong Programming experience **in Python, Scala, and core Kafka concepts**.
* Hands-on experience **building ETL pipelines**, Visualizations, Analytics based quality solutions in-house using AWS, Azure Databricks, and other Open-source frameworks.
* Extensive experience in working with various distributions of Hadoop like enterprise versions of **Cloudera, Hortonworks, and good knowledge of MAPR distribution and AWS EMR (Elastic Map Reduce).**
* Hands-on experience in **Amazon Web Services (AWS)** platform using **EMR**, **AWS Kinesis, AWS EMR, Aws Glue**, **AWS LAMBA,S3**, **EC2**, **Lambda**, **CloudWatch** and cloud computing using **AWSRedShift.**
* Deep understanding of Cloud Architectures including **AWS, Azure, GCP.**
* Setup and build AWS infrastructure various resources, **VPC, EC2, S3, AWS EMR, AWS Kinesis, IAM, EBS, Simple Notification Service (SNS), Simple Queue Service (SQS), Security Group, Auto Scaling, RDS** in **Cloud Formation JSON** templates.
* Expertise in relational database systems (RDBMS) such as **My SQL, Oracle, MS SQL, and No SQL database systems like HBase, MongoDB and Cassandra.**
* Experienced in implementing schedulers using **Oozie, Airflow, Crontab, and Shell scripts.**
* Good working experience in importing data using Sqoop from various sources like RDMS, Snowflake**, Teradata, Oracle to HDFS** and performing transformations on it using **Hive, Pig, and Spark.**
* I have good Experience like **Kanban** is a subsect of the **Agile methodology** and functions within the broader Agile mentality.
* Proficient in usage of tools like**Erwin** (**Data Modeler**, Model Mart, navigator), ER Studio, IBM Meta Data Workbench, Oracle data profiling tool, Informatica, Oracle Forms, Reports, SQL Plus, Toad, Crystal Reports.
* Good Experience writing Pyspark script using jupyer notebook.
* Logical **data modeling** focuses on capturing the business requirements and creating a conceptual representation of the data.
* Experience with Software development tools such **as JIRA, GIT,and SVN**.
* Worked on JIRA for defect/issues logging & tracking and documented all my work using **CONFLUENCE.**
* Having experience in developing storytellingdashboards, data analytics, designing reports with visualization solutions using Tableau Desktop and publishing on to the **Tableau Server**.
* Experience in Text Analytics, developing different ***Statistical Machine Learning*, Data Mining solutions to various business problems and generating data visualizations using R,** SAS and Python and creating dashboards using tools like **Qlik Sense and Tableau**.
* Created reports using visualizations such as Bar chart, Clustered Column Chart, Waterfall Chart, Gauge, Pie Chart, Tree map etc. in **Power B**I.
* Flexible working Operating Systems like **Unix/Linux (Centos, Red hat, Ubuntu) and Windows Environments**

 **Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Technologies** | HDFS, YARN, Map Reduce, Hive, Pig, Impala, Sqoop, Storm, Flume, Spark, Apache Kafka, Zookeeper, Ambari, Oozie, MongoDB, Cassandra, Mahout, Puppet, Avro, Parquet, Snappy, Falcon. |
| **NO SQL Databases** | Postgres, HBase, Cassandra, MongoDB, Amazon Dynamo DB, Redis |
| **Hadoop Distributions**  | Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR, and Apache. |
| **Languages** | Scala, Python, R, XML, XHTML, HTML, AJAX, CSS, SQL, PL/SQL, HiveQL, Unix, Shell Scripting  |
| **Source Code Control** | GitHub, CVS, SVN,  |
| **Cloud Computing Tools** | Amazon AWS, Microsoft Azure, GCP |
| **Databases** | Teradata Snowflake, Microsoft SQL Server, MySQL, DB2 |
| **DB languages** | MySQL, PL/SQL, PostgreSQL, R-SQL, T-SQL & Oracle |
| **Build Tools** | Jenkins, Maven, Ant, Log4j |
| **Business Intelligence Tools** | Tableau, Power BI |
| **Development Tools** | Eclipse, IntelliJ, Microsoft SQL Studio, NetBeans |
| **ETL Tools** | Talend, Pentaho, Informatica, Ab Initio, SSIS |
| **Development Methodologies** | Agile, Scrum, Waterfall, V model, Spiral, UML |

**Professional Experience:**

**Global Atlantic Financial Group, Indianapolis, IN August 2022 to Present**

**Role: SeniorGCP Data Engineer**

**Responsibilities:**

* Experience in building multiple Data pipelines, end-to-end ETL, and ELT processes for Data ingestion and transformation in **GCP** and coordinating tasks among the team.
* I aminvolved in transferring the data from an **on-premises Cloudera** platform to the Google Cloud Platform (GCP).majorly on-premises Cloudera data is being moved to a GCP bucket in various file formats, such as Parquet, Avro, and CSV.
* Design and implement various layers of **Data Lake, Design star schema in Big Query**.
* Understanding of creating **DATA PROC** cluster management and configuration using GCP.
* Created real-time data streaming pipelines using **GCP Pub/Sub** using Data flow.
* Design, Develop, and maintain scalable and efficient data pipelines on GCP using **Java**.
* Developed Spark programs to parse the raw data, populate staging tables, and store the refined data in partitioned tables in the Enterprise Data warehouse.
* Developed Streaming applications using PySpark to read from Kafka and persist the data NoSQL databases such as HBase and Cassandra.
* Implemented **PySpark Scripts using SparkSQL** to access hive tables into a spark for faster processing of data.
* Used cloud shell SDK in GCP to configure the **services Data Proc, Data flow, Cloud SQL, PUB SUB, Storage and Big Query**.
* Migrating an entire Oracle database **to BigQuery and usingpowerBI or GCP looker for reporting**.
* Build data pipelines in airflow in GCP for ETl-related jobs using different **GCP Cloud composer airflow operators.**
* Developed streaming and batch processing applications using **PySpark to ingest data from the various sources into HDFS Data Lake.**
* Worked on Big Data Hadoop cluster implementation and data integration in developing large-scale system software
* Developed DDLs and DMLs scripts in SQL and HQL for analytics applications in RDBMS and Hive.
* Used cloud shell SDK in **GCP to configure the services Data Proc, Storage, and BigQuery.**
* Used Spark SQL to read the Parquet data and Avro data loaded tables in hive to Spark using **Scala**.
* Written Hive UDFs to implement custom functions in the Hive for aggregations.
* Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems/mainframe and vice-versa loading data into HDFS.
* Monitoring **YARN** applications Troubleshoot and resolve cluster-related system problems.
* Created shell scripts to parameterize the Hive actions in **Oozie workflow and for scheduling the jobs.**
* Populated HDFS and Cassandra with **huge amounts of data using Apache Kafka**.
* Worked as a key role in a team of developing an initial prototype of a **NiFi big data pipeline**. This pipeline demonstrated an end-to-end scenario of data ingestion and processing.
* Using the NiFi tool to check whether a message reached the end system or not.
* Developed the custom processor for NiFi.
* Worked on **NoSQL Databases such as HBase** and integrated with **PySpark for processing and persisting real-time streaming.**
* Experience in **GCP Dataproc, GCS, Cloud functions, Big Query.**
* Used **Oozie Scheduler systems** to automate the pipeline workflow and orchestrate the map-reduce jobs that extract and Zookeeper for **providing coordinating services to the cluster**.
* Assessed existing **and EDW (enterprise data warehouse**) technologies and methods to ensure our EDW/BI architecture meets the needs of the business and dataprise and allows for business growth.
* Worked on Big Data Integration and Analytics based **on Hadoop, SOLR, Spark, Kafka, Storm, and web Methods technologies.**
* Design and develop data pipelines for integrated data Analytics using **Hive, Spark, Sqoop, and MySQL.**
* Guide the scrum team and organization on **agile/Scrum** practices.
* Facilitate **scrum** events, including sprint planning, Daily Stand-up, Sprint Review, and Sprint Retrospective.
* Proficiency in developing Spark applications and using **Apache Livy** for job submission and management.

**Environment: CDH5, Hortonworks, GKE, GCP Data Flow, Apache Airflow, Apache Beam, Big Query, Cloud, Data flow, Data Composer Data Proc, Cloud SQL , Apache Hadoop 2.6.0, HDFS, Java 8, Hive Sqoop, HBase 1.1.2, Oozie 4.1.0, Storm 0.9.3, YARN, NiFi, Cassandra, Zookeeper, Spark, Kafka, Oracle 11g, MySQL, Shell Script, EC2, Tomcat 8, Spring 3.2, Build Tool Gradle 2.2, Source Control GIT, Teradata SQL, Apache Livy.**

**Citrix, Fort Lauderdale, FL April 2019 to July 2022**

**Role: SR AWS Data Engineer**

**Responsibilities**:

* Implemented Responsible **AWS solutions using EC2, S3, RDS, EBS, Elastic Load Balancer, and Auto scaling groups, Optimized volumes and EC2 instances**.
* Wrote Terraform templates for AWS Infrastructure as a code to build staging, production environments & set up build &automation for Jenkins.
* Written Terraform scripts to automate AWS services which include ELB, CloudFront distribution, **RDS, EC2, database security groups, Route 53, VPC,IAM, EBS Lambda, ECS Fargate, API Gateway,Subnets, Security Groups**, **AWS GLUE, AWS Lambda,Auto Scaling,AWS Kinesis, AWS Cloud formation , AWS EMR, RDS&S3 Bucket** in **Cloud Formation JSON** templates and and converted existing AWS infrastructure to **AWS Lambda** deployed via Terraform and AWS CloudFormation.
* Setup and build AWS infrastructure various resources, **VPC EC2, S3, IAM, EBS, Security Group, Auto Scaling,AWS Kinesis, AWS EMR, RDS** in **Cloud Formation JSON** templates.
* Used Spark SQL to read the Parquet data and Avro data loaded tables in hive to Spark using **Scala**.
* Installed, configured multiple operating systems onsite and provisioned similar instances on **AWS cloud** and set up data in **AWS** using **S3bucket** and configuring instance backups to **S3 bucket**.
* Hands-on Experience working on Migrations from On-Premises to **AWS Cloud**, created cloud front for accelerated delivery of websites, APIs, and content by routing the traffic using **Route 53**.
* Import data using Sqoop to load data from **Teradata to HDFS on an AWS S3**.
* Write Hive queries for ad-hoc reporting to the business.
* Participated in weekly release meetings with Technology stakeholders to identify and mitigate potential risks associated with the releases.
* Configured **Elastic Load Balancers (ELB) with EC2 Auto scaling groups**.
* **Created Amazon VPC to create public-facing subnet for web servers with internet access**, and backend databases & application servers in a private-facing subnet with no Internet access.
* **Created AWS Launch configurations based on customized AMI** and use this launch configuration to configure auto scaling groups.
* Utilized Puppet for configuration management of hosted Instances within AWS Configuring and **Networking of Virtual Private Cloud (VPC**).
* **Utilized S3 bucket and Glacier for storage** and backup on AWS.
* Using Amazon Identity Access Management (IAM) tool created groups & permissions for users to work collaboratively.
* Implemented /setup continuous project build and deployment delivery process using Subversion, **Git, Jenkins, IIS, Tomcat**.
* Connected continuous integration system with **GIT version control repository and continually** build as the check-in's come from the developer.
* Knowledge in build tools Ant and **Maven** and writing build.xml and pom.xml respectively.
* Knowledge in authoring pom.xml files, performing releases with the Maven release plug-in and managing Maven repositories. Implemented **Maven** builds to automate JAR and WAR files.
* Designed and built deployment using ANT/ Shell scripting and automate the overall process using **Git and MAVEN** and involves in testing like **Junit**
* Implemented a Continuous Delivery frameworks using **Jenkins, Ansible/puppet, and Maven & Nexus in Linux environment.**
* Wrote Terraform, Cloud formation templates for **AWS Infrastructure as a code to build staging, production environments & set up build & automations for Jenkins.**
* Design, development and implementation of performant ETL pipelines using Python**API (pyspark) of Apache Spark on AWS EMR.**
* Worked on ETL tool like **Talend** and Talend Administration Management Console (TMC) to schedule data ingestion and data analytics job.
* Performed Code Reviews and responsible **for Design, Code, and Test signoff**.
* Assigning work to the team members and assisting them in development**, clarifyingdesign issues, and fixing the issues.**
* Guide the scrum team and organization on **agile/Scrum** practices.
* Facilitate **scrum** events, including sprint planning, Daily Stand-up, Sprint Review, and Sprint Retrospective.
* Implement and configure **SSO** to ensure secure and seamless authentication for users accessing data systems.

**Environment: AWS, S3, Jenkins, Spark-Core, Spark-Streaming, SQL, Jira, Lambda, ECS Fargate, API Gateway Ab Initio, Data Stage, Scala, AWS Kinesis, Apache Airflow, VPC, Python, Kafka, Hive, EC2, Elastic Search, Impala, Cassandra, Tableau, Talend, ETL, Linux**

**AAL (American Airlines), Fort Worth, TX May 2017 to March 2019**

**Role: Azure Data Engineer**

**Responsibilities:**

* Exploring with Spark improving the performance and optimization of the existing algorithms in Hadoop using **Spark Context, Spark SQL, Data Frame, and Spark Yarn.**
* Worked on **Azure BLOB andAzure Data Lake** storage and loading data into **Azure SQL Synapse analytics.**
* Experienced in managing **Azure Data Lakes (ADLS**) and **Data Lake Analytics** and an understanding of how to integrate with other Azure Services.
* WE are using Azure **Cosmos** account is the basic unit of global distribution and high availability.
* we are using globally distributing our data and throughput across multiple Azure regions, we can add or remove Azure regions from our **Azure Cosmos** at any time
* Azure SQL Database Managed Instance is a fully managed service. Microsoft handles tasks such as patching, maintenance, backups, and high availability
* Expertise in designing end-to-end data ingestion pipelines, encompassing data extraction, transformation, and loading (ETL) processes, using **Azure Databricks** as a central component.
* knowledge onintegrating **Azure Automation** with other Azure services.
* **Azure Purview** is an advanced data governance solution from Microsoft that allows businesses to store, manage and secure data
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of **Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.**
* Designing, planning and implementation for existing On-Premises applications to **Azure Cloud**, Create Cache Memory on **Windows Azure** to improve the performance of data transfer between SQL Azure and WCF services, created **Azure key-vault** to store all the credentials for **SQL databases** and API keys.
* Created hybrid IoT solutions for allowing IoT hub on **Azure Stack Hub**, build **IoT** solutions with reliable and secure communications between IoT devices and your on-premises solutions.
* They collaborate with the **DataStage** developers to define the data requirements, create data mappings, and ensure the integrity and consistency of the data throughout the ETL processes
* Dealt with Windows **Azure Networks**, Virtual Machines, Cloud Services, Resource Groups, Express Route, Traffic Manager, VPN, Load Balancing, Application Gateways, Auto-Scaling. Exposed Virtual Machines and Cloud services in the **VNets(Virtual Networks)** to the internet using **Azure External Load Balancer**.
* Imported data from various Sources transformed and loaded into Data Warehouse Targets using Open Studio, **Azure Storage** mounted on Azure VM, **Azure Databricks** cluster, **Azure data lake storage**, **Azure RDBMS** service, **Azure MySQL**, Veeva Vault
* Architected and built a real-time analytics platform in **AZURE** using Spark Structured Streaming, and **Databricks** (Apache Spark/Hive/Parquet).
* Worked on migration of data from On-Prem SQL server to **Azure Synapse analytics.**
* Worked on Azure BLOB and Data Lake storage and loading data into **Azure SQL Synapse analytics.**
* Azure App Service such ad Perfect for building, deploying, and scaling web apps.
* Azure Purview has three main purpose-built apps to enhance its capabilities such as Data Map, Data Catalog, and Data Insights.
* Experience in leveraging **Azure Automation** to manage and automate processes in Azure virtual machines (VMs), Azure SQL Database, Azure Storage, Azure Active Directory, or any other Azure resources.
* Proficient in usage of tools like Erwin (**Data Modeler**, Model Mart, navigator), ER Studio, IBM Meta Data Workbench, Oracle data profiling tool, Informatica, Oracle Forms, Reports, SQL Plus, Toad, Crystal Reports.
* Experienced in integrating Hadoop with **Kafka**, experienced in uploading Clickstream data from to **HDFS**.
* Use **Spark Streaming's API**, which resembles batch processing, but operates on micro-batches of data.
* Experienced in loading dataset into **Hive** for **ETL** (Extract, Transfer and Load) operation
* Involved in file movements between **HDFS and AWS S3** and extensively worked with S3 bucket in AWS and converted all Hadoop jobs to run in **EMR by configuring the cluster** according to the data size.
* Wrote Spark applications **for Data validation, cleansing, transformations and custom aggregations** and imported data from different sources into Spark RDD for processing and developed custom aggregate functions using Spark SQL and performed interactive querying
* Worked on data pipeline creation to convert incoming data to a common format, prepare data for analysis and visualization, migrate between databases, share data processing logic across web apps, batch jobs, and APIs, consume large XML, CSV, and fixed-width files and created data pipelines in Kafka to replace batch jobs with real-time data.
* Involved in **converting Hive/SQL queries into Spark Transformations using Spark**
* RDDs and Scala and involved in using SQOOP for importing and exporting data between RDBMS and HDFS.
* Collected data **using Spark Streaming from AWS S3 bucket** in near-real-time and performs necessary Transformations and Aggregations on the fly to build the common learner data model and persistence the data in HDFS.
* Created AWS Glue job for archiving data from Redshift tables to S3 (online to cold storage) as per data retention requirements and involved in managing S3 data layers and databases including Redshift and Postgres.
* Processed the web server logs by developing multi-hop flume agents by using Avro
* Sink and loaded into MongoDB for further analysis and worked on MongoDB NoSQL data modeling, tuning, disaster recovery and backup.
* Developed a **Python Script to load the CSV files into the S3 buckets and created**
* AWS S3 buckets, performed folder management in each bucket, managed logs and objects within each bucket
* Worked with different file formats **like JSON, AVRO and parquet and compression techniques** like snappy and developed python code for different tasks, dependencies, SLA watcher and time sensor for each job for workflow management and automation using Airflow tool.
* Developed shell scripts for dynamic partitions adding to hive stage table, **verifying JSON schema change of source files, and verifying duplicate files in source location.**
* Worked with importing metadata into Hive using Python and migrated **existing tables and applications to work on AWS cloud (S3).**
* Integrated Hadoop into traditional ETL, accelerating the extraction, transformation, and loading of massive structured and unstructured data.
* Involved with writing scripts **in Oracle, SQL Server, and Netezza databases** to extract data for reporting and analysis and worked in importing and cleansing of data from various sources like DB2, Oracle, and flat files onto SQL Server with high volume data
* Container management using Docker by writing **Docker files and set up the automated build** on Docker HUB and installing and configuring Kubernetes.
* Worked extensively with importing metadata into **Hive and migrated existing tables** and applications to work on Hive and AWS cloud and making the data available in Athena and Snowflake.
* Extensively used Stash **Git-Bucket for Code Control and Worked on AWS Components** such as Airflow, Elastic Map Reduce (EMR), Athena, and Snowflake.
* **Environment: Kafka, Impala, Spark, HDInsight, Azure Data Factory, Azure Stack Hub**, **Azure Event Hub,Blob Storage**, **Databricks,Apache Airflow, Azure Data Lake, Apache Beam, Data stage, Ab Initio, Cloud Shell, Tableau, Talend, Flink, SQL, Python, Hive, Spark SQL, MongoDB, TensorFlow, Jira.**

**Yana Software Private Limited Hyderabad, India October 2015 to February 2017**

**Role: Data Engineer**

**Responsibilities:**

* Involved in the implementation of the project went through several phases namely: data set analysis, preprocessing data set, user-generated data extraction, and modeling.
* Participated in Data Acquisition with the Data Engineer team to extract historical and real-time data by using Sqoop**, Pig, Flume, Hive, MapReduce, and HDFS**.
* Wrote user-defined functions (UDFs) in Hive to manipulate strings, dates, and other data.
* Performed Data Cleaning, features scaling, and features engineering using pandas, and NumPy packages in Python.
* Worked on production data fixes by creating and testing SQL scripts.
* Deep-dived into complex data sets to analyze trends **using Linear Regression, Logistic Regression, Decision Trees**
* Prepared reports using SQL and Excel to track the performance of websites and apps
* Visualized data using Tableau to highlight abstract information
* Applied clustering algorithms i.e., **Hierarchical, K-means using Scikit, and SciPy**.
* **Performed Data Collection, Data Cleaning, Data Visualization, and Feature Engineering** using Python libraries such as Pandas, NumPy, matplotlib, and Seaborn.
* Optimized SQL queries for transforming raw data into MySQL with Informatica to prepare structured data for machine learning.
* Used Tableau for data visualization and interactive statistical analysis.
* Worked with Business Analysts to understand the user requirements, layout, and look of the interactive dashboard.
* Used SSIS to create ETL packages to Validate, Extract, Transform, and Load data into Data Warehouse and Data Mart.
* The lifetime values were classified based on the RFM model by using an XGBoost classifier.
* Maintained and developed complex SQL queries, stored procedures, views, functions, and reports that meet customer requirements using **Microsoft SQL Server**
* Participated in Building Machine Learning using python

**Environment:Python, PL/SQL scripts, Oracle Apps, Excel, IBM SPSS, Tableau, Big Data, HDFS, Sqoop, Pig, Flume, Hive, Map Reduce, HDFS, SQL, Pandas, NumPy, Matplotlib, Seaborn, ETL, SSIS, SQL Server, Windows.**

**Creator Technologies Pvt Ltd Hyderabad, India June 2013 to September 2015**

**Role: Data Engineer**

**Responsibilities:**

* Collaborated with business user's/product owners/developers to contribute to the analysis of functional requirements.
* Implemented **Spark SQL queries** that combine hive queries with Python programmatic data manipulations supported by RDDs and data frames.
* Used Kafka Streams to Configure **Spark streaming** to get information and then store it in HDFS.
* Extract Real-time feed using **Spark Streaming and convert it to RDD** and process data in the form of Data Frame and save the data in HDFS.
* Developing Spark scripts, UDFS using **Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.**
* Installed and configured Hadoop MapReduce HDFS Developed multiple MapReduce jobs in java for data cleaning and preprocessing.
* Installed and configured Pig andalsowritten Pig Latin scripts. Write **MapReduce job using Pig Latin.**
* Worked on analyzing Hadoop clusters using different **big data analytic tools** including HBase database and Sqoop.
* Worked on importing and exporting data from Oracle, **and DB2 into HDFS and HIVE using Sqoop for analysis, visualization, and generating reports.**
* Creating and inserting data into Hive tables for dynamically inserting data into data tables using partitioning and bucketing for EDW tables and historical metrics.
* Experienced in handling large datasets using **Partitions, Spark in Memory capabilities**, Broadcasts in Spark, Effective & efficient Joins, Transformations, and others during the ingestion process itself.
* Created ETL packages with different data sources (**SQL Server, Oracle, Flat files, Excel, DB2**, and Teradata) and loaded the data into target tables by performing different kinds of transformations using SSIS.
* Designed and developed data integration programs in a **Hadoop environment with No SQL data store** Cassandra for data access and analysis.
* Created partitions, **bucketing across the state in Hive to handle structured data using Elastic search.**
* Performed Sqooping for various file transfers through the HBase tables for processing of data to several **No SQL DBs- Cassandra, Mongo DB**.

**Environment:Hadoop, MapReduce, HDFS, Hive, python, Kafka, HBase, Sqoop, No SQL, Spark 1.9, PL/SQL, Oracle, Cassandra, Mongo DB, ETL, MySQL.**

**Educational Qualifications:**Bachelor’s in computer science from JNTUH Hyderabad in the year 2013