**Venkatesh**

 

**Senior Data Engineer**

**PROFESSIONAL SUMMARY**

* **Data Engineer, have over 10+ years** of experience with demonstrated professional **working** experience in Big Data technologies such as **Hadoop, Spark, Hive, Kafka, as well as cloud platforms like AWS, and GCP.**
* Technologies like Ingestion, **Data Modelling, Querying, Processing, Analysis, and Implementing** Enterprise level Systems Spanning **Big Data and Data Integration**
* Experience in Creating and executing Data Pipelines **in GCP, AWS.**
* Hands-on Experience on **Hadoop Distrib**ution **Platforms** Namely **IBM** **Big Insights, Hortonworks, Cloudera, and** Cloud **platforms GCP, AWS.**
* Expertise in **Big Data** **Technologies** and Hadoop Ecosystems such as **Spark, Spark-Scala, HDFS, Apache Flume**, **Hive, Sqoop, PIG, Spark-SQL, Kafka, Hue, Yarn and EPIC data sources.**
* Hands-on experience in Building Data pipelines and **Data** **marts** using the **Hadoop stack.**
* Experience in migrating existing on - premise databases to both Azure and Google Cloud Environments for better reporting experience.
* Experienced in building modern data warehouses in both **Azure Cloud** and **Google Cloud**, including building reporting in PowerBI or Data Studio.
* Experience working on Azure Services like **Data Lake, Data Lake Analytics**,**SQL Database, Synapse, Data Bricks, Data factory, Logic Apps** and **SQL Data warehouse** and GCP services Like **Big Query, Dataproc, Pub sub etc**.
* Highly experienced in using multi cloud services to achieve maximum efficiency and performance for data application, used Azure and Google cloud platform for building extensive data pipelines for doing terra scale of data analysis.
* Good experience in writing Spark applications. In-depth understanding of Spark Architecture including Spark Core, Spark SQL, Data Frames, Spark Streaming, Kafka streaming.
* Hands on experience in building CI/CD Azure pipelines using the Microsoft agent and running python file related tests.
* Hands-on experience in using Amazon Web Services including **EC2, S3, EMR, Lambda,** Kinesis, Redshift, Step function, AWS Glue, Athena, Boto3 and CloudFormation.
* Software development involving cloud computing platforms like Amazon Web Services (AWS), and Google Cloud (GCP).
* Worked on **AWS EC2, EMR and** S3 to create clusters and manage data using S3.
* Strong understanding of the entire AWS Product and Service suite primarily **EC2, S3, VPC, Lambda,** Redshift, Spectrum, Athena, EMR(Hadoop) and other monitoring service of products and their applicable use cases, best practices and implementation, and support considerations.
* Experience on Hadoop Distribution Platforms: Hortonworks**, IBM Big Insights** and Cloudera and Cloudera platforms GCP and AWS.
* Worked on Data migration from On-premises Microsoft SQL Server to Amazon Redshift and ETL re-engineering of SSIS packages to Spark SQL/PySpark code using AWS EMR.
* Install and configure Apache Airflow for AWS S3 bucket and created dags to run the Airflow.
* Automated Regular AWS tasks like snapshots creation using Python scripts.
* Designed data warehouses on platforms such as **AWS Redshift, AWS SQL Data Warehouse, and other high-performance platforms.**
* Hands on expertise with AWS Databases such as RDS(Aurora), Redshift, DynamoDB and Elastic Cache (Memcached & Redis)
* Hands on experience working Amazon Web Services (AWS) using Elastic Map Reduce (EMR), Redshift, and EC2 for data processing. ELT - Snowflake- streams and tasks, Snow pipe, S3, Snow SQL, Python, Five Tran, DBT
* In-depth understanding of **Apache Spark job** execution components like **DAG, Executors, Task Scheduler**, Stages, and Spark Steaming.
* Hands-on Experience in **GCP, Big query, cloud functions, data proc**.
* Strong Experience in **Control-M Job Scheduler Tool, Apache Airflow, ESP, D-series**. Monitored the jobs on a call base to close Incident tickets.
* Hands-on experience in **GCP, Big Query, GCS, Cloud Composer, Cloud functions, Cloud dataflow, Pub/Sub**, **cloud shell, GSUTIL, bq command**-line utilities, Data Proc
* Hands-on experience with **Amazon EC2, S3, RDS, IAM, Auto Scaling, CloudWatch, SNS, Athena, Glue, Kinesis, Lambda, EMR, Redshift, DynamoDB, and other services of the AWS family**.
* Expertise in using the **CICD JENKINS** pipeline to deploy the codes into production.
* Designed and developed the programming paradigm to support data collection and filtering processes in the data warehouse and Hadoop data mart.
* Proficient in SQL databases **MSSQL Server, MySQL (RDBMS), Oracle DB, Postgres, and MongoDB.**
* Worked on **multiple proje**cts using **Informatica developer** tool **IDQ Latest Versions.**
* Ability to **adapt quickl**y to evolving technology, a strong sense of responsibility and accomplishment.
* Strong **Hadoop and platform support** experience with the entire suite of tools **and services in major Hadoop** Distributions - **Cloudera, Amazon EMR, and Hortonworks.**
* Hands-on experience in working **Agile environment** and following release management, **Golden rules.**
* Experience in Version control tools such as **GIT** and **Urban Code Deployment** **(UCD) tools.**

# CERTIFICATION DETAILS:

* Coursera Microsoft Certified: Azure Data Engineer Associate.
* Certified Google Cloud Data Engineer.

# EDUCATIONAL DETAILS:

* Master’s in computer information system (M.Sc.), New England University, New Hampshire, NH, 2023
* Bachelor’s in computer science, St. Mary’s Engineering College, India, 2013

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Tools** | Hadoop, Hive, Apache Spark, PySpark, HBase, Kafka, YARN, Sqoop, Impala, Oozie, Pig, Map Reduce, Zookeeper and Flume |
| **Hadoop Distributions** | EMR, Cloudera, Hortonworks. |
| **Cloud AWS** | EC2, S3, EMR, RDS, Glue, Presto, Lambda, RedShift |
| **GCP** | Cloud Storage, Big Query, Compute Engine, Cloud Composer, Data Proc, Data Flow, Pub-sub, GCP Looker |
| **AZURE** | Azure (Data lake, Data factory, Databricks), Designing/modifying data warehouses, Programming Scala, Python, MSBI (SSIS, SSAS, SSRS), Data Visualization, Spark SQL, Sql Server, Microsoft Power BI, Microsoft Office, GCP(Big query, Dataproc, Pub sub) |
| **BI and Data Visualizations** | GCP Looker, ETL -Informatica, SSIS, Talend, Tableau, and Power BI |
| **Relational Databases** | Oracle, SQL Server, Teradata, MySQL, PostgreSQL, and Netezza |
| **No SQL Databases** | Cassandra, MongoDB, and HBase |
| **Programming Languages** | Scala, Python and R, Java |
| **Scripting** | Python and Shell scripting |
| **Build Tools** | Apache Maven and SBT, Jenkins, Bitbucket |
| **Version Controls** | GIT and SVN |
| **Operating Systems** | Unix, Linux, Mac OS, CentOS, Ubuntu, and Windows |
| **Tools** | PUTTY, Putty-Gen, Eclipse, IntelliJ, and Toad |

**PROFESSIONAL EXPERIENCE**

**Client: Comcast, PA OCT 2023- Present**

**Roles: Sr. Data engineer**

**Responsibilities:**

* Used Apache airflow in the **GCP** Composer environment to build data pipelines and explored various airflow operations like bash operator, Hadoop operators, and branching operators.
* Used cloud shell **SDK in GCP** to configure the services **Data Proc, Storage, and big query.**
* Extensive use of cloud shell **SDK in GCP** to configure/deploy the services using **GCP Big Query.**
* Developed **ELT processes** from the files from Abinitio and Google Sheets in **GCP** with **Compute being data proc** (**PySpark) and big query.**
* Changing the existing **Data Models** using Erwin for enhancements to the existing Data ware house projects.
* Worked on **Google Cloud Platform (GCP) services like Compute Engine, cloud load balancing, cloud storage, and cloud SQL.**
* Develop and deploy the outcome using **Spark and Scala** code in the Hadoop cluster running on **GCP.**
* Experience in **Google Cloud components**, **Google container builders, and GCP client libraries.**
* Worked on the design and development of Informatica mappings, and workflows to load data into the staging area, data warehouse, and data **marts in SQL Server and Oracle.**
* Using **Server** **Manager**, created batches and sessions to transport data at present intervals and on- demand.
* Performed data transformations for **ML OPs, including adding calc**ulated columns, maintaining relationships, establishing various metrics, merging, and appending queries, changing values, splitting columns, and grouping by **Date** **and Time Column.**
* Design, implement, and maintain infrastructure components using Terraform to ensure scalability, security, and reliability on Google Cloud Platform (**GCP**).
* Used Apache Airflow 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.
* ister and optimize data storage solutions, including **Google Cloud Storage (GCS)** for object storage, Cloud SQL for relational databases, Bigtable for NoSQL databases, and Firestore for scalable document databases.
* Implement data lifecycle management strategies, including backup, replication, and archival policies to ensure data availability, integrity, and compliance.
* Use Helm charts to package, version, and deploy Kubernetes applications and microservices on Google Kubernetes **Engine (GKE) and other Kubernetes clusters on GCP**.
* Worked with EMR to transform and move big data into AWS data stores and databases in S3 and DynamoDB.
* Created and launched AWS EC2 instances to **execute jobs on EMR to** store the results in S3.
* Designed and setup Enterprise Data Lake to provide support in multiple areas such as Analytics, processing, storing and Reporting of big and rapidly changing data.
* Designed and implemented big data solutions on **AWS**, using Spark Streaming for real-time data processing and analysis, and Spark SQL for querying and transforming large datasets stored in S3.
* Developed an ETL pipeline using AWS Glue to extract, transform, and load data from various sources into AWS S3 buckets.
* Implemented end-to-end batch data processing **using Spark SQL on the** **AWS** cloud platform.
* Monitored infrastructure performance and ensured reliability through **AWS** Cloudwatch.
* Automated the deployment process of the data pipeline **using CI/CD** with **AWS** Code Pipeline.
* Utilized AWS CloudFormation for deploying services, enhancing consistency and reducing errors.
* Worked with Hadoop distribute file system (HDFS), S3 Storage, big data formats like parquet, JSON.
* Configured AWS Redshift clusters, AWS Redshift spectrums for querying, and AWS Redshift Data share for transferring the data among clusters.
* Involved in Designing, Architecting and **implementing scalable cloud-based web applications** using **AWS** and GCP.
* Created a real-time chat application with Java for the server-side components using WebSockets and Spring Boot.
* Implemented a real-time data processing pipeline with **GCP Pub/Sub** and **Dataflow**, achieving low-latency data ingestion and transformation.
* **Architected a microservices-based application** on GCP, utilizing Google Kubernetes Engine (GKE) for orchestration and management of containerized services.
* Designed a cost-effective backup and disaster recovery strategy using GCP Cloud Storage and Cloud SQL, ensuring data availability and resilience.
* Migrated on-premises infrastructure to **GCP**, leveraging Compute Engine and **Cloud VPN** to seamlessly integrate with existing systems.
* Implement algorithms to anonymize PHI within healthcare datasets while preserving data utility for analysis.
* Developed a machine learning model on GCP AI Platform, using AutoML to automate model training and hyperparameter tuning.
* Implemented an end-to-end CI/CD pipeline on **GCP using Cloud Build**, Cloud Source Repositories, and Cloud Run for continuous integration and deployment.
* Designed a data lake architecture on GCP using Cloud Storage and **BigQuery**, enabling unified data storage and analytics.
* Maintain Docker images and Docker files to standardize development and production environments, ensuring reproducibility and minimizing configuration drift in data workflows.
* Optimize Docker container configurations and resource allocation **(e.g., CPU, memory**) for data-intensive workloads to achieve optimal performance and cost efficiency on GCP.
* Utilize Docker logging mechanisms and integrate with monitoring tools (e.g., **Prometheus, Stackdriver**) to monitor containerized applications and troubleshoot performance issues on GCP.
* Worked closely with the SME to get an understanding of the business requirements.
* Built an IoT data processing pipeline with Python for data collection from sensors and Java for backend processing and storage.
* Implement CI/CD pipelines using **Google Cloud Build** and **Kubernetes** Engine for automated deployment and scaling of data processing and analytics workflows.
* Developed a financial trading platform using **Java** for the trading engine and transaction management.
* Configured and managed a secure GCP environment using Identitay and Access Management (IAM) policies and VPC Service Controls.
* Used PySpark and Spark to implement data quality checks, data transformation, and data validation processes.
* Implemented a real-time analytics platform using **GCP** **BigQuery** and Data Studio, providing interactive dashboards and reports for business insights.
* Developed serverless applications on **GCP** using Cloud Functions and Cloud Firestore, reducing infrastructure management overhead.
* Implemented a data governance framework on **GCP using Data Catalog** and Cloud DLP, ensuring data compliance and security.
* Python-based analytics module to predict patient no-shows and optimize scheduling.
* Performed data quality checks to ensure high quality of data.
* Working on data validation and data profiling to ensure the accuracy of the data between the warehouse and source systems.

**Environment:** Python, CSS and HTML, XML, JSON, X, MySQL, Postgre SQL, EC2, SQS, SNS, IAM, S3, and DynamoDB Data-Integration, Data Mapping, Data Profiling, Data Lake, PySpark, Spark, AWS, Hadoop, GCP, Spark, AWS Kinesis, Parquet, JSON, SNS, Redshift, Apache Airflow, MYSQL, EC2, S3, AWS Athena, and Glue.

**7-Eleven, Dallas, Tx. Aug 2022 – Sept 2023**

**Azure/GCP Data Engineer**

**Responsibilities:**

* Hands-on major components in Hadoop Echo Systems like Spark, **HDFS, HIVE, HBase, Zookeeper, Sqoop, and** **Oozie.**
* Conducting **Exploratory data ana**lysis in Jupiter notebooks **using Python libraries** and sharing the data analysis
* Developing **Sqoop** jobs to ingest data from various systems of records into **Enterprise Data Lake.**
* Development of Spark jobs in **PySpark and Spark SQL** to run on top of **hive** **tables** and create transformed data sets for downstream consumption.
* Proficient in **Data** **Processing**, with expertise in creating clusters using **Apache Hadoop and Apache Spark** on **Google Cloud Platform** (GCP).
* We are using a Data flow automatic template for managing services **GCS TO BIG Query and BIG query to GCS.**
* Implemented real-time message subscription using **Google Cloud Pub/Sub**, facilitating seamless data processing for large datasets during on-premises to **Google Cloud Platform migration.**
* Worked on Google Cloud Platform (**GCP) in all the big data products Big Query, Cloud Data Proc, Google Cloud Storage, and Composer (Air Flow as a service).**
* Extensive knowledge and understanding of **ETL and ELT** for real-time batch processing and streaming processes using **GCP**.
* Have good experience with **GCP** tools such as **BIG QUERY, GCS COMPOSER, GCS, PUB/SUB, DATA FLOW, DATA PROC, Cloud SQL, GKE, and And Airflow.**
* We are collecting data from source systems like **IBM Data** streams such as IoT devices and sensors, Text files, spreadsheets, images, video, and audio recordings, and finally data is moving to GCP Cloud
* Explore GCP's data processing and streaming services, such as **Google Cloud Dataflow** or **Apache Beam, which** are like **IBM** Streams and can be used to process real-time data.
* Ingesting data from various source systems like **Oracle, SQL Server, Flat files, and JSONs.**
* Performance tuning spark and hive jobs by reading execution plans**, DAGs, and Yarn logs.**
* Creating generic shell scripts to submit Hadoop and spark jobs on **EMR** and on-prem edge nodes.
* Writing Complex **Spark SQL** codes to clean, join, transform, and aggregate the datasets and publish them for the Power BI team to produce operational scorecards.
* Writing custom Python modules for reusable **Python code.**
* Working on designing the MapReduce **and Yarn workflow** and writing MapReduce scripts, performance tuning, and debugging.
* Design new applications for high transaction processing & scalability to seamlessly support future modifications and growing volume of data processed in environment.
* Implement solutions to run effectively in cloud and improve the performance of big data processing and high volume of data being handled by the system to provide better customer support.
* Work with business process managers and be a subject matter expert for transforming vast amounts of data and creating business intelligence reports using the state-of-the-art big data technologies (Hive, Spark, Scoop, and NIFI for ingestion of big data, python/bash scripting /Apache Airflow for scheduling jobs in GCP/Google’s cloud-based environments).
* Migrated an **Oracle SQL ETL** to run on google cloud platform using cloud **dataproc & bigquery, cloud pub/sub** for **triggering the airflow jobs.**
* Worked on using presto, hive, and spark-sql, big query using python client libraries and building interoperable and faster programs for analytics platforms.
* Hands on experience in using all the big data related services in Google Cloud Platform.
* Used apache airflow in **GCP composer environment to build** data pipelines and used various airflow operators like bash operator, Hadoop operators and python callable and branching operators.
* Moved **Data between big** query and **Azure Data Warehouse using ADF and create** Cubes on **AAS** with lots of complex **DAX language** for memory optimization for reporting.
* Built reports for monitoring data loads into **GCP** and drive reliability at the site level
* Created multiple proof-of-concepts using **PySpark** and deployed them on the Yarn cluster, comparing Spark's performance to that of **Hive** and **SQL/Teradata**.
* Integrated Teradata Warehouse into the **EMR cluster**. **Developed** **shell** **scripts** to load data from the
* Teradata Staging area to the Teradata data mart. Handled **Errors & and** tuned performance in Teradata queries and utilities.
* Set up databases in **GCP using RDS**, storage using **S3 bucket, and configuring** instance backups to S3 bucket. Prototype **CI/CD** system with **GitLab on GKE** utilizing **Kubernetes** and D**ocker** for the runtime environment for the **CI**/**CD** systems to build test and deploy.
* Solving appropriate partition and **bucketing schemes and making** sure correct load policies are employed so data can be stored as **per requirements.**
* Generates **ETL scripts to transform, flatten**, and enrich the data from source to target **using AWS Glue a**nd created event-driven **ETL pipelines with AWS Glue**.
* Documenting **data flow diagrams** and technical logic in confluence.
* Worked on setting up **AWS DMS and SNS** for data transfer and replication and used **SQL on the new AWS** Databases like **RedShift** and **Relation Data Services.**
* Optimizing and tuning the Redshift environment, enabling queries to perform up to **100x faster for Tableau** and **SAS Visual Analytics.**
* Creating Oozie workflows, **Coordinators, and scheduling handshake jobs in Control-M.**
* Created Session Beans and Controller Servlets for handling **HTTP requests from Talend**.
* Using **Control-M for orchestration**, the workflow will be executed through Airflow, , and Glue to run the jobs and glue catalog to capture the **metadata, Lambda to process** the source **data freshness check** and **capture** the data **quality metrics.**
* Working with governance teams to ensure metadata management, **data lineage, and technical metadata** are correctly **updated** for each **data asset.**
* Working with production support teams and administration teams to **ensure** correct access controls are set up on each **hive** **database**.
* Working with the master-feature branch model and committing the code with appropriate comments.
* Attending sprint **planning**, **agile** ceremonies, and demoing the work products on a bi**-weekly basis.**

**Environment:** , S3, Athena, Hive, Impala, Lambda, Python, PySpark, Spark SQL, Oracle 11g/12c, Teradata, Jira, Bitbucket, Power BI, Control-M, Airflow, Talend, CI/CD pipelines, GKE, Redshift, SAS.

**AGL -INDIA OCT 2020 – FEB 2022**

**Big Data Engineer**

**Responsibilities:**

* Responsible for building the data lake in **Amazon AWS**, ingesting structured shipment and master data from the **AWS API Gateway, Lambda, and Kinesis Firehose** into **s3 buckets.**
* Implemented **Data** pipelines for big data processing using **Spark transformations** and **Python API** and clusters in **AWS**
* Create complex **Sql queries in Teradata Data** **Warehouse environment** to test the **data flow across** all the stages
* Integrated data sources from **Kafka** (**Producer and Consumer API) for data stream-processing in Spark using AWS Network.**
* Extensively involved in designing the **SSIS** packages to load data into **Data Warehouse**
* Built customer insights on customer/**service utilization, bookings & CRM** data using Gainsight.
* Executed process improvements in data workflows using Alteryx processing engine and **SQL**
* Collaborated with business owners of products for understanding business needs and automated business processes and data **storytelling in Tableau**
* Implemented **Agile Methodology for building** the **data applications** and framework development
* Implemented business processing models using predictive & **prescriptive analytics on** transactional data with regression.
* Implemented **Logistic**, Random forests **ML models with Python** packages to **decide insurance purchase** by a **Confidential member**
* Responsible for Designing and Creating **SSAS Cubes from the Data Ware House**.
* Developed data processing **pipelines (processes 40-50 GB daily) using Python libraries** with Google internal tools such as **Pantheon** **ETL, Plx scripts with SQL**
* Automated feature engineering mechasims using **Python scripts** and deployed on **Google cloud platform** (G**CP) and BigQuery.**
* Pulling the data from **data lake (HDFS**) and massaging the data with various **RDD transformations.**
* Prepared Technical design based on Functional requirements and modified Spark scripts and resolved Production bugs on various scripts for data transformations in Python.
* Extensive expertise in Data Warehousing on different database (s), as well as data modeling, both logical and physical data modeling tools like **Erwin, Power Designer and ER Studio.**
* Parsed JSON and log data and designed the data flows using Apache **NiFi - Processors, Funnels**
* Build **Tableau** and **Data studio dashboards** based on Marketing campaign requirements and presented them to Sales Directors
* Built a data lake as a cloud based solution in **AWS** using **Apache Spark and** provide visualization of the **ETL** orchestration using **CDAP tool.**
* Implemented the project development using **Agile** processes in Kanban boards and 2**-week sprints**
* Developed **machine learning** **models** such as **Random** forests **using** **TensorFlow**
* Prepared Data models and schema on GCP for different projects based on star and snowflake schema designs

**Environment:** Python, SQL, Tableau, Bigdata, Datalake, Alteryx, Hive, CRM, OLAP, Excel, Data Robot.

**Client: Schlumberger, INDIA JUN 2017– Sept 2020**

**Role: Hadoop developer/Big Data Engineer**

**Responsibilities:**

* Work closely with **Business Analysts** and **Product** Owners to understand the requirements.
* Used Joins in **SPARK** for making smaller datasets to large datasets without shuffling data across nodes.
* Developed **Spark** **Streaming** jobs using Python to read messages **from Kafka.**
* Write **Scala** **program** for **spark** **transformation** in **Dataproc**.
* Used **Spark Streaming** to receive real-time data from Kafka and store the stream data to HDFS using Python and **NoSQL** databases such as **HBase and Cassandra**
* Prototyped analysis and joining of customer data using **Spark in Scala and processed it to HDFS**
* Implemented Spark in **EMR** for **processing Big Data** across our **One** **Lake** **in AWS System**
* Consumed and processed data from **DB2.**
* Developed Spark/Scala, Python for regular expression (regex) project in the **Hadoop/Hive** environment with **Linux/Windows** for big data resources.
* Downloaded **JSON files from AWS S3 buckets.**
* Implemented **ETL using AWS RedShift/Glue.**
* Developed applications using Spark to implement various aggregation and transformation functions **of Spark** **RDD and Spark SQL.**
* Used **Scala** components to implement the credit line policy based on the conditions applied on **spark data** **frames.**
* Data sources are extracted, transformed, and loaded to generate **CSV data files with Python programming** and **SQL queries.**
* Using the g-cloud function with Python to load Data into Big query for arrival CSV files in the GCS bucket.
* Write a program to download a **SQL** **Dump** from their equipment maintenance site and then load it **in the GCS** bucket.
* On the other **side load this SQL dump** from the **GCS bucket to MYSQL (hosted in Google Cloud SQL)** and load the Data from **MYSQL to Big query using Python, Scala, spark, and Data proc.**
* Process and load bound and unbound Data from Google pub/subtopic to big query using cloud Dataflow with Python.
* Create firewall rules to access **Google Data** proc from other machines.

**Environment:** Spark, Hive, ETL, Data Stage, Informatica, Talend, AWS S3, EMR, RedShift, EC2, Python, SQL, Nifi, Airflow, Jupiter, Cloudera, Hadoop, Apache Kafka, Apache flume.

**Client: Royal Dutch shell, Hyderabad, India June 2014 – May 2017**

**Role: ETL/Informatica**

**Responsibilities:**

* Involved in the evaluation of functional and non-functional requirements.
* Installed and configured Hadoop MapReduce **HDFS Developed multiple** **MapReduce jobs** in **java** for data cleaning and pre-processing.
* Designed end-to-end **ETL** solutions, involving the creation of Informatica workflows and mappings using **Informatica Power Center 10.X** **and Informatica Cloud IICS**.
* Created and maintained **SQL-based ETL (Extract, Transform, Load**) processes to **extract, clean and load data** into the **data warehouse.**
* Developed SSIS packages to **extract, Transform and Load (ETL) data** into the **data warehouse database** from heterogeneous **databases/data sources.**
* Installed and configured **Pig** and also written **Pig Latin scripts.**
* Wrote **MapReduce j**ob **using** **Pig Latin.**
* Involved in managing and reviewing **Hadoop log files.**
* Imported data using Sqoop to load data from **MySQL to HDFS** on **regular basis.**
* Developing **Scripts** and **Batch** Job to schedule various Hadoop Program.
* Written **Hive queries** for data analysis to meet the business requirements and **created hive tables** and worked on them **using Hive** **QL**.
* Importing and exporting data into **HDFS** and **Hive** **using Sqoop.**
* Experienced in defining **job flows**.
* Got good experience with **NOSQL** database **SOLR** HBase.
* Involved with putting data into Hive tables and constructing hive queries that will execute internally in a map reduce fashion.
* Created a custom **Filesystems** plug-in for Hadoop that allows it to access files on the Data Platform.
* This plugin **enables** **Hadoop** MapReduce applications **HBase Pig** and **Hive** to function normally and directly access files.
* Designed and implemented **MapReduce-based large**-scale parallel **relation-learning system**.
* Extracted feeds form social media sites such as **Facebook Twitter using Python scripts**.
* For internal purposes, I have set up benchmarked **Hadoop clusters**.

**Environment**: Hadoop, MapReduce, HDFS Hive, Java Hadoop distribution of Horton Works, Cloudera, Pig, HBase, Linux, XML, MySQL, MySQL Workbench Java 6, Eclipse, Oracle 10g PL/SQL, SQL PLUS Sub Version Cassandra.