**Ganta**

**Sr. Data Engineer/Cloud Engineer**

**Professional Summary:**

* Overall, 11+ years of experience in IT industry which around 5+ years of experience in Big Data in implementing complete Hadoop solutions.
* Working experience in developing applications involving Big Data technologies like **Map Reduce**, **HDFS, Hive, Sqoop, Pig, Oozie, HBase, Nifi, Airflow, Spark, Scala, Kafka and Zookeeper and ETL(DataStage**).
* Extensively worked on **Spark** using **Scala** on cluster for computational (analytics), installed it on top of Hadoop performed advanced analytical application by making use of Spark with Hive and SQL/Oracle.
* Experience in migrating the data using **Sqoop** from HDFS to Relational Database System and vice-versa according to client's requirement.
* Working knowledge of Amazon Elastic Cloud Compute (**EC2**) infrastructure for computational tasks and Simple Storage Service (**S3**) as Storage mechanism.
* Running of Apache Hadoop, CDH and Map-R distributions, Elastic Map Reduce (**EMR**) on (**EC2**).
* Expertise in developing Pig Latin scripts and **Hive Query Language** (HQL).
* Extensive expertise using the core Spark APIs and processing data on an **EMR** cluster.
* Experience in Developing **Spark applications** using Spark - SQL in Azure Databricks for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns
* Development and maintenance of **Scala applications** that are executed on the **Cloudera** platform
* Used various Project Management services like JIRA for tracking issues, **GitHub** for various code Reviews and worked on various version control tools like CVS, GIT, and SVN.
* Extensive knowledge on Spark Core APIs, **Data Frames, Spark-SQL**.
* Very good understanding of **partitioning** and **bucketing** in Hive.
* Good knowledge of cloud integration with AWS using Elastic Map Reduce (**EMR**), Simple Storage Service (**S3**), EC2, Redshift.
* Deployed, managed, and operated a scalable, highly available, and fault tolerant ETL pipelines on **AWS** ecosystem.
* Hands on experience in implementing an DevOps pipeline using **Jenkins** and the AWS CI / CD tools.
* Experience with **pyspark** for using spark libraries by using python scripting for data analysis.
* Designed HIVE queries to perform data analysis, data transfer and table design.
* Hands on experience on Data Analytics Services such as **AWS Athena**, **AWS Glue Data Catalog**.
* Experienced in writing Ad-Hoc queries using Cloudera Impala also used Impala analytical functions.
* Designed ETL workflows on **Tableau**, Deployed data from various sources to **HDFS**.
* Aggregation from multiple file formats including **Parquet, Avro, XML, JSON, CSV,** ORCFILE and other compressed file formats Codecs like Zip, Snappy.
* Configured and maintained **Jenkins** to implement the CI process and integrated the tool with Ant and Maven to schedule the builds.
* Hands-on knowledge in core Java concepts like **Exceptions, Collections, Data-structures, I/O. Multi-threading, Serialization and deserialization** of streaming applications.
* Experience with use-case development, with Software methodologies like **Agile** and **Waterfall**.
* Proven ability to manage all stages of project development Strong Problem Solving and Analytical skills and abilities to make Balanced and Independent Decisions.

**Technical Skills:**

**Languages**: Python, SQL, shell script, Java, Scala, HTML

**Big Data Technologies:** HDFS, Hive, Pig, Spark, Kafka, Sqoop, Flume

**Cloud Technologies**: AWS (S3,EC2, EMR,GLUE,ATHENA,REDSHIFT,IAM), CI/CD, Nifi, Databricks, Snowflake, Azure,GCP(Dataproc, Big Query)

**No Sql Databases**: HBase, Cassandra

**IDE’s & Tools**: At scale, Eclipse, IntelliJ Idea, putty, Visual Studio

**Databases**: My Sql, Oracle, Teradata

**Operating Systems:** Unix, windows, Linux

**Business Intelligence Tools** : Tableau,ThoughSpot

**Education**

* Bachelor’s in Electrical Engineering, Osmania University, Hyderabad—June 2013
* Master’s in Electrical Engineering, Gannon University, PA --- May 2015

**Professional Experience:**

**Client: JPMC July 22 – till date**

**Location: NJ**

**Title: Sr Big Data Engineer**

Aladdin Wealth™ is an industry-leading platform powering the future of wealth management. Built on the portfolio and risk analysis technology used by BlackRock and sophisticated institutions globally and transforming wealth businesses by changing how advisors engage with individual investors. This application and interface from JPMC with Blackrock is managed by the Data Engineer team.

* Collaborated with Business Clients (Blackrock Team) to understand requirements and provided a clean data creating with highly efficient optimized **ETL pipeline** for transparent and secure models for risk analysis on **AWS** Ecosystem.
* Collected, cleansed and provided modelling and analysis of structured and unstructured data used for business initiatives which resulted in saving with Millions dollar risk management factors.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the **AWS EMR Spark cluster**.
* Worked on supporting production team using **AWS, Tableau, Jenkins, CD/CI pipelines.**
* Created schema in AWS connects with performance optimization using bucketing & partitioning.
* Integrated **Athena** with **AWS Glue Data Catalog**, allowing to create a unified metadata repository across various services, crawl data sources to discover schemas and populate Catalog with new and modified table and partition definitions, and maintained schema versioning.
* Optimize the spark jobs to run on **Kubernetes** **Cluster** for faster data processing.
* Loaded data incrementally and optimized **Parquet** files write with **AWS** Glue
* scheduled and orchestrated data pipelines with AWS Glue workflows
* Enhanced the existing pipelines with **Spark-optimizations** and hive optimizations and worked in explicitly on **pyspark**.
* Worked on data lake Ingestion on AWS(S3) of RAW files and Transformation in Curated and Created Views/ tables for analytics.
* Consumed and normalized different file formats like **JSON,XML,AVRO,ORC,Parquet**.
* Developed Security Framework to provide fine grained access to objects in AWS S3 using **Lambda**, **DynamoDb**.
* Implemented end to end architecture & implementation assessment of various AWS services like **S3, EMR, Redshift, Athena,AWS Glue.**
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, like files from **AWS** **S3** to **AWS** **Redshift** **Table**.
* Sourced data from multiple sources like HDFS and **API** into spark RDD and applied business logics using **pyspark** to generate the output responses.
* Used Amazon **Athena** to run interactive query service that makes it easy to **analyze** **data** in Amazon **S3**.
* Created impactful dashboards on **Excel** and **Tableau** for data reporting, extracted interpreted and analyzed data to identify key metrics and transform data into meaningful, actionable information.

**Environment**: AWS(Athena, EMR,S3,EC2,IAM,Redshift,Glue..etc),PySpark, Spark SQL, HDFS, Bash script, SQL, Tableau,Kubernetes, Shell scripting, Linux, Databricks, Bitbucket .

**Client: Credit Suisse Financial July 21 – Jun 22**

**Location: New York City, NY**

**Title: Sr Big Data Engineer**

As a Big Data Engineer with F1-Core Team, worked on enabling the transformation of the Finance functions, delivering the holistic business solutions to support sustained evolution of Credit Suisse’s Finance capability front and back. F1- MDL is a layer which fills and operate as a single stop for all data provisioning capabilities across deliveries.

* Implemented the complete big data pipeline with **real-time processing** and **batch processing**
* Maintained the models in the HDFS and made the predictions for the live streaming data from the sensors and stored the predicted values for visualization purpose.
* Implemented multiple use cases with **spark core, spark SQL, spark streaming and spark ML**.
* Translate business requirements into maintainable software components and understand impact
* Created **Hive-SQL** scripts for reconciliation component in F1 MDL layer as the source raw data and the refined data are all maintained with proper validations and counts before opening to consumer platform.
* Worked with **Python**, to develop analytical jobs using **PySpark** API of spark.
* Implemented Spark using Scala and utilizing Data frames and **Spark** **SQL** **API** for faster processing of data.
* Developed multiple POCs using **PySpark** and deployed on the YARN cluster, compared the performance of Spark, with Hive
* Involved in developing a linear regression model to predict a continuous measurement for improving the observation on wind turbine data developed using spark with Scala API.
* Worked on reading and writing multiple data formats like **JSON**, ORC, Parquet on HDFS using Scala.
* Extensively worked on **Sqooping** process from Teradata to Data-lake.
* Performed querying of both managed and **external** **tables** created by Hive.
* Maintained the Log and Audit information in Hive/SQL tables and provided support with error fining and error handling.
* Developed Spark code in **Scala** using Spark SQL and performed **transformations**.
* Worked ETL phases of the data like data cleansing, data massaging and data clean-up, filtering the data which is useful for model building.
* Created Partitioned **Hive tables** and worked on them using **HiveQL.**
* Loading Data into HBase using **Bulk Load and Non-bulk load**.
* Worked on continuous Integration tools **Jenkins and automated** jar files at end of day.
* Used HiveQL to analyze the partitioned and **bucketed** **data**, Executed Hive queries on **Parquet tables** Stored in Hive to perform data analysis to meet the business specification logic.
* Designed columnar families in Cassandra and Ingested data from **RDBMS**, performed data
* Worked on data lake Ingestion of RAW files and Transformation in Curated and Created Views/ tables in Hive for analytics.
* Create/Modify shell scripts for **scheduling** various data **cleansing** scripts and ETL load process.
* Involved in the performance tuning of the model by verifying the model results time to time and worked with different attributes and features for getting the better performance and in model rebuilding activities.
* Worked with Offshore team to build new and enhance the existing ETL workflows with highly optimized techniques with spark (computations) and hive (data retrievals)
* Configured workflow engine using Apache **Airflow and Shell Scripting** for managing and scheduling Hadoop jobs.
* Scripted an **ETL Pipeline** with Scala that ingests files from **AWS** **S3** to **AWS** **Redshift** **Table**
* Implemented and Installed the setup of multi node Cloudera cluster on **AWS** cloud.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the **AWS EMR Spark cluster**.
* Used **Tableau** for creating dashboards for the BI reports and visualization.
* Developed testing scripts in Python and prepare test procedures, analyze test results data and suggest improvements of the system and software.

**Environment:** Hadoop, Spark 2.0, HIVE, spark, Python, Scala, Impala, bash script, Spark SQL, HDFS, MapReduce, Apache, Yarn, SQL, Spark 2.0, Pyspark, Scala, AWS Tableau, AWS(Athena, EMR,S3,EC2,IAM,Redshift tables), Autosys, Kafka, Data warehouse, GIT bash, Jenkins, Maven.

**Client: Thrivent Financial Oct 20 – June 21**

**Location: Minneapolis, MN**

**Title: Sr Big Data Engineer**

As a Big Data Engineer in the Risk Management Team where the Bank wanted to store, process & manage the huge amount of data in day-to-day operations collected from various sources. The system majorly checks the credibility of the customer & looks for the credit risks.

* Configured and Monitored Apache Airflow with DAGs to run the **Airflow**.
* Worked with **Cloudera** 5.12.x and its different components.
* Used **Spark** 2.1.x API to stream data from various sources in real-time
* Developed Spark code in **Scala** using Spark SQL and performed **transformations**.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Created Custom workflows, custom screens and custom fields on JIRA.
* Used **Spark** and **Spark**-**SQL** to read the **parquet** **data** and create the tables in hive using the **Scala** **API**.
* Development of cubes involving multiple facts and dimensions.
* Development of calculations, leveraging Query Datasets.
* Experience in using **Avro, Parquet, RCFile and JSON file** formats, developed UDF in Hive and Pig.
* Developed Spark code using Python and Spark-SQL for faster processing and testing.
* Implemented **Spark** **RDD** **transformations** to Map business analysis and apply actions on top of transformations and performed validations.
* Created AWS connect application to consume real-time data from **Kafka** sources and applied real-time data analysis models that we can update on new data in the stream as it arrives.
* Athena queries data directly in Amazon S3. Used Amazon Athena with ANSI SQL support and works with a variety of standard data formats, including **CSV, JSON, ORC, Avro, and Parquet**.
* Perform analysis, coding, testing and development of **Apache Kafka, AngulatJS, J2EE and ArcGIS** Components.
* Configure Project Environment for Flex - Spring Java Communication using Blaze DS Remoting.
* Experience with **pyspark** for using spark libraries by using python scripting for data analysis.
* Worked on Migrating Teradata objects into the Snow Flake environment.
* Used **Tableau** for creating dashboards for the BI reports and visualization.
* Written adhoc SQL scripts and ETL in snowflake and worked with the client while validating.
* Optimized the hive query using hive optimization techniques.
* Configured workflow engine using Apache **Airflow** for managing and scheduling Hadoop jobs.
* Developed build & deployment scripts using **MAVEN**.
* Worked on Kafka to collect and load the data on Hadoop file systems.

**Environment:** Spark 2.0, Scala, AWS, Hadoop, HIVE, **Airflow**, Pyspark, Python, Scala, Impala, bash script, Spark SQL, HDFS, MapReduce, Apache, Yarn, SQL, Machine learning, Node, Tableau, Angular, Athena, Autosys, Kafka, Data warehouse, EMR, Hive, GIT bash, Jenkins, Maven.

**Client: BCBS Of Texas Jul 2019- Sep 20**

**Location: Richardson TX**

**Title: Sr. Big Data/Cloud Engineer**

As an Sr. Big Data Engineer of Data Lake team at BCBS of Texas, I have created ETL Pipelines using Spark, Scala and created hive tables for end users. The FEP Life with Blue project provides the ability to easily identify the FEP Life with Blue membership population and provide risk stratification data. The data will be sourced from the Data Lake and loaded to Teradata and will be used to create extracts for Aerial. The data will be used to manage the risk and stratification of the FEP Life with Blue membership population.

**Responsibilities:**

* Developed Spark Applications by using Scala and Implemented Apache Spark data processing Project to handle data from various RDBMS and Streaming sources.
* Worked on reading and writing multiple data formats like **JSON**, ORC, Parquet on HDFS using Scala.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, 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.
* Used Scala sbt to develop Scala coded spark projects and executed using spark-submit.
* 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**, write-back tool and backwards.
* Configured Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS.
* Experience designing **DAGs** using **AirFlow**.
* Developed Pre-processing 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.
* Installed and configured apache airflow for workflow management and created workflows.
* Extensively worked on **Sqooping** process from Teradata to Data-lake.
* Involved in Migrating Teradata objects into Snow Flake environment.
* Worked on data lake Ingestion of RAW files and Transformation in Curated and Created Views/ tables in Hive for analytics.
* Performed data quality issue analysis using snow SQL by building analytical warehouse on **snowflake**.
* Configured workflow engine using Apache **Airflow and Shell Scripting** for managing and scheduling Hadoop jobs.
* Involved in developing a linear regression model to predict a continuous measurement for improving the observation on wind turbine data developed using spark with Scala API.
* Develop ETL Process using SPARK, SCALA, Python, HIVE and HBASE.
* Experienced in quiring the **Snowflake**.
* Implemented Spark using **Scala** and **Spark** **SQL** for faster testing and processing of data.
* Implemented Spark using Scala and utilizing Data frames and **Spark** **SQL** **API** for faster processing of data.
* Good understanding of Cassandra architecture, replication strategy, gossip, snitches etc.
* Designed columnar families in Cassandra and Ingested data from RDBMS, performed data
* Developed multiple POCs using **PySpark** and deployed on the YARN cluster, compared the performance of Spark, with Hive
* Worked from Scratch in Configurations of Kafka such as Mangers and Brokers
* Maintained the Log and Audit information in Hive/SQL tables and provided support with error fining and error handling.
* Tuned complex Java, Scala, Spark, **Airflow** jobs
* Used HiveQL to analyze the partitioned and bucketed data, Executed Hive queries on Parquet tables Stored in Hive to perform data analysis to meet the business specification logic.
* Experience in using Avro, Parquet, RCFile and JSON file formats, developed UDF in Hive and Pig.
* Developed Sqoop and Kafka Jobs to load data from RDBMS, External Systems into HDFS and HIVE.
* Developed Oozie coordinators to schedule **Pig** and **Hive** scripts to create Data pipelines.
* Written several Map reduce Jobs using Java API, also Used Jenkins for Continuous integration.
* Setting up and worked on Kerberos authentication principals to establish secure network communication on cluster and testing of HDFS, Hive, Pig and Map Reduce to access cluster for new users.
* Used **Databricks** **XML** plug-in to parse the incoming data in the XML format, and generate the required XML as output
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.

**Environment**: Spark, Spark-Streaming, Spark SQL, **Airflow**, Azure, Spark (2.1), MapR, HDFS, Hive, Pig, Apache Kafka, Sqoop, Java (JDK SE 6, 7), Scala, Shell scripting, Linux, Databricks, MySQL Oracle Enterprise DB, SOLR, Jenkins, IntelliJ, Oracle, Git, Oozie, Tableau, MySQL, Soap, Cassandra and Agile Methodologies.

**Client: Tracfone Jul 18 – Jun 2019**

**Location: Miami, FL**

**Title: Sr. Big Data Engineer**

The Project was developed to provide visibility to Sales/Returns Revenue from E-commerce Channel and POSA Retailers as well as improve visibility/accuracy of Dealer/Agent & Retailer Channel Gross Adds, which will enable Smarter Revenue and Retention Initiatives.

**Responsibilities:**

* Writing Spark Job reading sensor data from **Kafka,** implementing machine learning model and persisting the data with prediction in **Cassandra** Database.
* Worked on building Spark (which has multiple Spark Sources and Spark Sinks) Workflow which helps data analysts to run the predictive maintenance workflow.
* Extensively worked with **Kafka** for reading and producing the data from the different sources like IoT devices, sensor and OBD devices.
* Maintained the models in the HDFS and made the predictions for the live streaming data from the sensors and stored the predicted values for visualization purpose.
* Implemented multiple use cases with **spark core, spark SQL, spark streaming and spark ML**.
* Translate business requirements into maintainable software components and understand impact
* Makes sure that quality standards are defined and met.
* Optimize the spark jobs to run on **Kubernetes** **Cluster** for faster data processing
* Migrated **Azure** **platform** to AWS Platform as per Business Requirement.
* Implemented process builder Spring Services, Controllers to trigger Spark-Submit (start & stop job) through REST call.
* Developed Spark Programs using **Scala** and Java API's and performed transformations and actions on RDD's.
* Worked on ELK stack for log monitoring for Analysing the spark logs and different sensor devices logs.
* Worked on Spark using **Scala** on cluster for computational (analytics), installed it on top of Hadoop performed advanced analytical application by making use of Spark with Hive and SQL/Oracle.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD, **Scala** and Python.
* Extensively used Text mate on Ubuntu for writing Java, **Scala** and shell scripts.
* Migrating on-prem **ETLs** from MS SQL server to Azure Cloud using Azure Data Factory and Databricks
* Develop ETL Process using SPARK, **SCALA**, HIVE and HBASE.
* Provide guidance to development team working on **Scala** as ETL platform.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the Cloud Dataflow on **GCP**
* Worked ETL phases of the data like data cleansing, data massaging and data clean-up, filtering the data which is useful for model building.
* Worked and learned a great deal from **AWS** cloud services like **EC2, S3, EBS, RDS and VPC**.
* Implemented Elastic Search on Hive data warehouse platform.
* Developed spark applications in **Scala** on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
* Worked on reading and writing multiple data formats like **JSON, ORC, Parquet on HDFS** using **Scala**
* Worked with ELASTIC MAPREDUCE and setup Hadoop environment in **AWS EC2 Instances**.
* Involved in different phases of model building activities like verifying the performance of a model by comparing the different features of a problem statement.
* Heavily involved in testing Snowflake to understand best possible way to use the **cloud** **resources**
* Involved in **Amazon EMR cluster** management
* Deployed the spark jobs and monitor their status in the cluster and analyzing the **EMR** **logs** after the job processing done.
* Good knowledge on **Microsoft Azure ML studio** and done multiple POCs to the clients for different problem statements.
* Good knowledge on the AWS s3 file system and stored the models in the s3 buckets and done the predictions from there.
* Good knowledge on some of the visualization tools like **Kibana** and **tableau** reporting.

**Environment**: Jdk1.8, Spring with REST API, My SQL, Hadoop 2.x, HDFS, Map Reduce, Azure Databricks, Airflow, Spark core, spark SQL, Spark Streaming, Scala, Spark ML lib, Cassandra, Elastic Search, ELK stack, Bit Bucket, Gradle, Oozie.

**Client: DaVita Inc, Jul 16 – Jun 18**

**Location: Nashville, TN**

**Title: Hadoop/Spark Developer**

The project name is Reggie, which is an upstream for all the applications and deals with the patient registrations at the clinic or hospitals. When a patient comes to a receptionist the patient is enrolled in Reggie database. If the patient were new, there were would be a new enrolment with health Insurance and other required details and if the Patient were an existing customer and any changes in their Health care Insurance then there would be an update. Through this data, the doctor is assigned, and patient is prescribed.

**Responsibilities:**

* Worked on migrating **Map Reduce** programs into Spark transformations using Spark and Scala.
* Developed Spark jobs using Scala on top of **Yarn** for interactive and Batch Analysis.
* Experienced in querying data using Spark SQL on top of Spark engine for faster data sets processing.
* Written java code to format **XML documents**, uploaded them to solve server for indexing.
* Optimized **Hive QL Scripts** by using execution engine like **Tez.**
* Worked on **Ad hoc queries**, Indexing, Replication, Load balancing, and Aggregation in **Mongo DB**.
* Processed the **Web server logs** by developing Multi-hop flume agents by using Avro Sink and loaded into **Mongo DB** for further analysis, also extracted files from **Mongo DB** through **Flume and processed.**
* Using PIG to do transformations, event joins, filter boot traffic and some Pre-Aggregations before storing the data onto **Azure Database.**
* Developed Spark Programs using **Scala** and Java API's and performed transformations and actions on RDD's.
* Expert knowledge on **Mongo DB, NoSQL** **data modelling**, tuning, and **disaster recovery backup** used it for distributed storage and processing using **CRUD.**
* Extracted and restructured the data into Mongo DB using import and export command line utility tool.
* 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**.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on **GCP**.
* Experience in setting up Fan-out **workflow** in flume to design v shaped architecture to take data from many sources and ingest into **single sink.**
* Experience in creating tables, dropping and altered at run time without blocking updates and queries using **HBase and Hive.**
* Experience in working with different join patterns and implemented both **Map and Reduce Side Joins**.
* Involved in Migrating Teradata objects into **SnowFlake** environment.
* Experience in implementing Spark RDD's in **Scala**.
* Wrote **Flume configuration** files for importing streaming log data into **HBase with Flume**.
* Imported several transactional logs from web servers with Flume to ingest the data into **HDFS**. Using Flume and Spool directory for loading the data from local system (**LFS) to HDFS**.
* Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in **Scala**
* Developed **Map Reduce** programs in Java for parsing the raw data and populating **staging Tables.**
* Experience in setting up the whole app stack, setup and debug log stash to send **Apache logs to AWS Elastic search.**
* Used Zookeeper to coordinate the servers in **clusters** and to maintain the data consistency.
* Experienced knowledge over designing Restful services using **java-based APIs like JERSEY**.
* Used **OOZIE Operational Services** for batch processing and scheduling **workflows dynamically.**
* Supported in setting up **QA environment** and updating configurations for **implementing scripts with Pig,**

**Environment:** HDP 2.3, Hadoop, HDFS, Hive, Map Reduce, Azure, GCP, SOLR, Impala, MySQL, Oracle, Sqoop, Flume, Spark, SQL Talend, Python, Yarn, Pig, Oozie, Linux-Ubuntu, Scala, Ab Initio, Tableau, Maven, Jenkins, Java (JDK 1.6), Cloudera, JUnit, agile methodologies

**Client: Phillips**

**Location: Monroeville, TX**

**Title: Hadoop Developer Jul 15–Jun 16**

This project was about to give the users a detailed ETL flow of number of products which are shipped to the dealer Agents and returned.

**Responsibilities:**

* Responsible for data extraction and data ingestion from different data sources into Hadoop Data Lake by creating **ETL** **pipelines** using Pig and Hive.
* Responsible for importing data to HDFS using Sqoop from different **RDBMS** servers and exporting data using Sqoop to the RDBMS servers after aggregations for other ETL operations.
* Implemented Partitioning, **Dynamic** Partitions and Buckets in HIVE for efficient data access and manages the users.
* Create/Modify shell scripts for **scheduling** various data **cleansing** scripts and ETL load process.
* Developed testing scripts in Python and prepare test procedures, analyze test results data and suggest improvements of the system and software.
* Used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
* Created mappings using pushdown optimization to achieve good performance in loading data into **Netezza**.
* Worked on creating FTP jobs and SFTP connection setup between multiple UNIX servers for receiving and sending files as part of project work.
* Maintained and supported systems for data warehouse and business intelligence (BI) initiatives, including software, servers, and networks.
* Worked with Database Administrators, Business Analysts and **Cognos** **Report** **Developers** to conduct design reviews and validate the developed models.
* Unit tested the code with real time data to make sure the required functionality is achieved.
* Optimized the SQL and jobs for best performance.

**Environment**: Hive, Pig, Data stage, HDFS, Cognos Reporting, DB2, UNIX, Oracle, Teradata, Autosys, DB2, SQL server, Mainframes.

**Client: Genpact,**

**Location: Hyderabad**

**Title: ETL consultant Jun 12 – Dec 13**

**Responsibilities:**

* Used the DataStage Designer to develop processes for **Extracting, Cleansing, Transforming**, **Integrating, and Loading** data into Data warehouse.
* Requirement analysis and gathering to provide technical and architectural support to the team.
* Documented the Purpose of Mapping to facilitate the personnel to understand the process and incorporate the changes as and when necessary.
* Primarily involved in Job Design, Technical Reviews and Troubleshooting of jobs. Extensively involved in different Team review meetings and conferences with remote team.
* Developed various **Server and Parallel jobs** using Oracle, ODBC, FTP, Peek, Aggregator, Filter, Funnel, Copy, Hash File, **Change Capture, Merge, look up, Join, Sort, Merge, Lookup** stages.
* Developed PL/SQL Procedures, Functions, Packages, Triggers, Normal and Materialized Views.
* Worked with Reporting team for extensively reporting using **Data mart for Slice & Dice, Drill Down and Drill through.**
* Defect Tracking, unit testing, defect reporting, analyzing results and documentation.
* Designed **Data Stage parallel jobs** using designer to extract data from various source systems, Transformation and conversion of data, Load data-to-data warehouse and Send data from warehouse to third party systems like Mainframe.
* Extensively used Processing Stage like **Lookup** **Stage** to perform lookup operations based on the various target tables, modify stage to alter the record schema of the input data set, Funnel Stage to combine various datasets into a single large dataset and Switch stage to **trigger** the required output based on a specific condition.
* Designed jobs Using complex stages like **Transformer, Change Capture, Change Apply, Remove Duplicates, Join, Lookup, Merge, Funnel and Aggregator**
* Performed ETL Performance tuning to increase the ETL process speed.
* Have addressed production, UAT issues, proper action was taken accordingly based on priority and requirement
* Developed Scripts to transfer flat file between two different servers.
* Developed Operational Manual Documentation.
* Associated with other team workers for implementation of data stage best practices for better performance.

**Environment: Data stage 7.5.1, oracle 9i, Toad for Oracle**