**Tajuddin Mohammad**

**Email:** **tajdataengineer@gmail.com**

**Contact: 832 653 9103**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Data Engineer with over all 9 years of experience in building data intensive applications, tackling challenging architectural and scalability problems. Worked with Big Data stack and cloud services to build robust ETL/ELT solutions.

**PROFESSIONAL SUMMARY:**

* Experienced in IT Project Management activities including project scoping, estimation, planning, finalization of technical/ functional specifications, resource optimization & quality management of product/ software application.
* Experienced in data science techniques, like data pre-processing, feature engineering, model evaluation etc.
* Worked on various business problems for automobile, banking, telecom & Insurance domain using Machine Learning techniques.
* Have created data processing pipeline and Machine Learning model on Spark using Python and SparkML.
* Experienced in Big Data stack; HDFS, Spark, Sqoop and Hive. Proficient in RDBMS, SQL & Unix.
* Implement ETL process in Alteryx to extract data from multiple sources (SQL Server, XML, Excel, CSV) and schedule workflows.
* Interpreted data, analyze results using statistical techniques and provide ongoing reports.
* Create alteryx workflows with advanced analytics
* Experienced in ETL, BI, metadata modelling & data visualization,
* Good understanding of Data Lake, Cloud based data stacks (AWS, Databricks) and DWH concepts.
* An innovator keeps looking for opportunities to apply programming or data science for problem solving.
* Worked with parquet, avro file formats.
* Have worked in different roles, Independently with client & user.
* Experience in moving data between GCP and Azure using Azure Data Factory.
* Create clusters in Google Cloud and manage the clusters using Kubernetes(k8s). Using Jenkins to

deploy code to Google Cloud, create new namespaces, creating docker images and pushing them to

container registry of Google Cloud.

*  Excellent interpersonal and communication skills, experienced in working with senior level

managers, business people and developers across multiple disciplines.

* Strong problem solving, analytical and have the ability to work both independently and as a team.
* Highly enthusiastic, self-motivated and rapidly assimilate with new concepts and technologies.
* Experience with ETL workflow Management tools like Apache Airflow and have significant

experience in writing the python scripts to implement the workflow.

* Experience in identifying Bottlenecks in ETL Processes and Performance tuning of the production

applications using Database Tuning, Partitioning, Index Usage, Aggregate Tables, Session

partitioning, Load strategies, commit intervals and transformation tuning

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| ETL | Informatica Power Center 10.x/9.6/9.1, AWS Glue, Talend 5.6, SQL Server Integration Services (SSIS) |
| Databases & Tools | MS SQL Server 2014/2012/2008, Teradata 15/14, Oracle 11g10g, SQL Assistant, Erwin 8/9, ER Studio |
| Cloud Environment | Snowflake, AWS RDS, AWS Aurora, Redshift, EC2, EMR, S3, Lambda, Glue, Data Pipeline, Athena, Data Migration Services, SQS, SNS, ELB, VPC, EBS,RDS, Route53, Cloud Watch, AWS Auto Scaling, Git, AWS CLI, Jenkins, Microsoft Azure, Google Cloud Platform(GCP), GCS , Bigquey and DataProc |
| Reporting Tools | Tableau, Power BI |
| Big Data Ecosystem | HDFS, Map Reduce, Hive/Impala, Pig, Sqoop, Hbase, Spark, Scala, Kafka. |
| Programming languages | Unix Shell Scripting, SQL, PL/SQL, Perl, Python, T-SQL |
| Data Warehousing & BI | Star Schema, Snowflake schema, Facts and Dimensions tables, SAS, SSIS, and Splunk |

**EDUCATIONAL BACKGROUND:**

Masters in Business Analytics(Minor in Data science and Cyber security), TAMUC, 2020

Bachelor of Electronics and Communications Engineering , JNTUK, 2012

**MANAGEMENT SKILLS:**

Agile Development

Project and Team Management

JIRA, Trello, Git, Excel, PowerPoint

**PROFESSIONAL EXPERIENCE:**

**Data Engineer | April 2021 - Present**

**Client: FORD | Location: MI (Remote)**

**Responsibilities:**

* Data Ingestion of transactional data from Hive data warehouse, and flat files into staging area
* Used panda’s data frame to read and clean the data from data warehouse and flat files.
* Automating and scheduling the data integration of flat files using Oozie scheduler.
* Converting pandas to Pyspark for distributed programming to achieve the objectives of the business.
* Create and place alteryx server on MS Azure VM
* Used Pyspark data frames to read and transform data according to the business needs.
* Preparing the data in the format for the Qlik view downstream teams to consume.
* Used Alluxio, Pandas and Pyspark to read data from HDFS paths
* Dealt with software updates of the vehicle especially MMOTA.
* Extract transform and load data from source systems to GCS , spark sql and Dataproc.
* Ingested the delta lake file format into data lake service 2 and then based on the partitions.
* Extract the flat files form Azure blob storage and create snapshot tables to store the data based on each run.
* Using catalog to check the existing tables.
* Develop end to end pipeline using databricks by entering the parameter widgets.
* Setup the interactive cluster from the databrick and scheduling the job, monitoring the workload of the pipelines in ganglia databricks.
* Creating visualization for the software updates that needs to be pushed to the ford vehicles.
* Manage end to end complex data migration, conversion, and data modeling (using Alteryx, SQL), and create visualization using tableau to develop high quality dashboards.
* Worked on a Hive module, created Hive Tables and loaded data to analyze it using Qlik dashboards.
* Assisted Analytics team in identifying new business opportunities through mailing list.
* Assisted Client Analytics with the implementation of solutions based on the integration of a variety of data sources.
* Create analytical application in alteryx designer and store on alteryx server for non - technical user
* Assisted the Media team in organizing tasks and schedules in carrying out a substantial business analysis by using internal ticketing systems.
* Migration to GCP using services cloud storage and Big query to process data.
* Monitoring Bigquery, Dataproc and cloud Data flow jobs.
* Open SSH tunnel to Google DataProc to access yarn marnager to monitor spark jobs.
* Submit spark jobs using gsutil and spark submission get it executed in Dataproc cluster.
* Worked with Tekton for Continuous Integration and Continuous Deployment of data pipeline.
* Worked with Snowflake to integrate third part application tools to process and transform the data in snowsql and store in data warehouse for analysis.

**Environment:** Pyspark,GCP, Snowflake, python, SQL, Alteryx, Qlik, HDFS, Hadoop Oozie, Data mining, Tekton CI/CD

**Data Engineer | Jan 2020– Feb 2021**

**Company: Marvel Technology Solutions Inc (Contracting) | Location: Dallas, TX**

 **Responsibilities:**

* Involved in full Software Development Life Cycle (SDLC) - Business Requirements Analysis,

preparation of Technical Design documents, Data Analysis, Logical and Physical database design,

Coding, Testing, Implementing, and deploying to business users.

* Developed complex mappings using Informatica Power Center Designer to transform and load the

data from various source systems like Oracle, Teradata, and Sybase into the final target database.

*  Analyzed source data coming different sources like SQL Server tables, XML files and Flat files then

transformed according to business rules using Informatica and loaded the data in to target tables.

* Designed and developed a number of complex mappings using various transformations like Source

Qualifier, Aggregator, Router, Joiner, Union, Expression, Lookup, Filter, Update Strategy, Stored

Procedure, Sequence Generator, etc.

* Involved in creating the Tables in Greenplum and loading the data through Alteryx for Global Audit

Tracker.

* Analyzed large and critical datasets using HDFS, HBase, Hive, HQL, PIG, Sqoop and Zookeeper.
* Data Extraction, aggregations, and consolidation of Adobe data within AWS Glue using PySpark.
* Developed Python scripts to automate the ETL process using Apache Airflow and CRON scripts in

the UNIX operating system as well.

* Worked on Google Cloud Platform (GCP) services like compute engine, cloud load balancing, cloud

storage and cloud SQL.

* Developed data engineering and ETL python scripts for ingestion pipelines which run on AWS

infrastructure setup of EMR, S3, Glue and Lambda.

* Changing the existing Data Models using Erwin for Enhancements to the existing Datawarehouse

projects.

* Used Talend connectors integrated to Redshift – BI Development for multiple technical projects

running in parallel.

* Performed Query Optimization with the help of explain plans, collect statistics, Primary and

Secondary indexes. Used volatile table and derived queries for breaking up complex queries into

simpler queries. Streamlined the scripts and shell scripts migration process on the UNIX box.

* Using g-cloud function with Python to load Data in to Big query for on arrival csv files in GCS bucket.
* Created iterative macro in Alteryx to send Json request and download Json response from

webservice and analyze the response data.

* Migrated data from Transactional source systems to Redshift data warehouse using spark and AWS

EMR.

* Experience in Google Cloud components, Google container builders and GCP client libraries.
* Supported various business teams with Data Mining and Reporting by writing complex SQLs using

Basic and Advanced SQL including OLAP functions like Ranking, partitioning and windowing

functions, Etc.

* Expertise in writing scripts for Data Extraction, Transformation and Loading of data from legacy

systems to target data warehouse using BTEQ, FastLoad, MultiLoad, and Tpump.

* Worked with EMR, S3 and EC2 services in AWS cloud and Migrating servers, databases, and

applications from on premise to AWS.

* Tuning SQL queries using Explain analyzing the data distribution among AMPs and index usage,

collect statistics, definition of indexes, revision of correlated sub queries, usage of Hash functions,

etc…

* Developed shell scripts for job automation, which will generate the log file for every job.
* Extensively used spark SQL and Data frames API in building spark applications.
* Written complex SQLs using joins, sub queries and correlated sub queries. Expertise in SQL Queries

for cross verification of data.

* Extensively worked on performance tuning of Informatica and IDQ mappings.
* Creating, maintain, support, repair, customizing System &amp; Splunk applications, search queries and

dashboards.

* Experience on data profiling &amp; various data quality rules development using Informatica Data

Quality (IDQ).

* Create new UNIX scripts to automate and to handle different file processing, editing and execution
* Experience in cloud versioning technologies like GitHub.
* Integrate Collibra with Data Lake using Collibra connect API.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources

like S3, ORC/Parquet/Text Files into AWS Redshift.

* Create firewall rules to access Google Data proc from other machines.
* Write Scala program for spark transformation in Dataproc.
* Providing technical support and guidance to the offshore team to address complex business

problems.

* Environment: Informatica Power Center 9.5, AWS Glue, Talend, Google Cloud Platform(GCP),

PostgreSQL Server, Python, Oracle,

Environment: AWS, Talend, Postgresql, Oracle, ERWIN, Github, EMR, cron, Unix shell scripting

**Data Engineer Intern | September 2019 – December 2019**

**Company: CAG | Location: Richardson TX**

Responsibilities:

* Data migration from sql server to Azure cloud
* Involved in gathering business requirements, logical modelling, physical database design, data sourcing, data transformation and loading.
* Ingesting data from oracle and web API to
* Performed data analytics usning pyspark on databrick platform.
* Worked on azure ADF and moving the data from salesforce data warehouse.

Environment: Azure Blob Storage, HDInsight, Oracle, ERWIN, GitHub, ADF, Azure Data Lake, cron, Unix shell scripting

 **Sr.Data Engineer| April 2016- Nov 2017**

**Company: Capgemini | Client: Talk Talk | Location: India**

**Responsibilities:**

* Involved in gathering business requirements, logical modeling, physical database design, data

sourcing and data transformation, data loading, SQL, and performance tuning.

* Used SSIS to populate data from various data sources, creating packages for different data loading

operations for applications.

* Created various types of reports such as complex drill-down reports, drill through reports,

parameterized reports, matrix reports, Sub reports, non-parameterized reports and charts using

reporting services based on relational and OLAP databases.

* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction,

transformation, and aggregation from multiple file formats.

* Extracted data from various sources like SQL Server 2016, CSV, Microsoft Excel and Text file from

Client servers.

* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to

AWS Redshift.

* Built S3 buckets and managed policies for S3 buckets and used S3 bucket and Glacier for storage

and backup on AWS.

* Developed Spark scripts using Python on AWS EMR for Data Aggregation, Validation and Adhoc

querying.

* Performed data analytics on DataLake using pyspark on databricks platform
* Involved in creation/review of functional requirement specifications and supporting documents for business systems, experience in database design process and data modeling process.
* Implementation and delivery of MSBI platform solutions to develop and deploy ETL, analytical,

reporting and scorecard / dashboards on SQL Server using SSIS, SSRS.

* Extensively worked with SSIS tool suite, designed and created mapping using various SSIS

transformations like OLEDB command, Conditional Split, Lookup, Aggregator, Multicast and Derived

Column.

* Scheduled and executed SSIS Packages using SQL Server Agent and Development of automated

daily, weekly and monthly system maintenance tasks such as database backup, Database Integrity

verification, indexing and statistics updates.

* Developed and executed a migration strategy to move Data Warehouse from Greenplum to Oracle

platform.

* Load data into Amazon Redshift and use AWS Cloud Watch to collect and monitor AWS RDS

instances within Confidential.

* Worked extensively on SQL, PL/SQL, and UNIX shell scripting.
* Expertise in creating PL/ SQL Procedures, Functions, Triggers and cursors.
* Loading data in No SQL database (HBase, Cassandra)
* Expert level knowledge of complex SQL using Teradata functions, macros and stored procedures.
* Developing under scrum methodology and in a CI/CD environment using Jenkin.
* Developed UNIX shell scripts to run batch jobs in Autosys and loads into production.
* Do participate in architecture council for database architecture recommendation
* Utilized Unix Shell Scripts for adding the header to the flat file targets.
* Used Teradata utilities fastload, multiload, tpump to load data.
* Preparation of the Test Cases and involvement in Unit Testing and System Integration Testing.
* Deep analysis on SQL execution plan and recommend hints or restructure or introduce index or

materialized view for better performance

* Deploy EC2 instances for oracle database.

**Environment:** Pyspark, Spark sql, Tableau, MS Excel, Amazon EMR/Redshift, Hadoop, databricks

**Data Analyst / Engineer| Dec 2012- March 2016**

**Company: Capgemini | Client: GE Capital| Location: India**

**Responsibilities:**

* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import,

and Data Export.

* Worked on Design, Development and Documentation of the ETL strategy to populate the data from

the various source systems using Talend ETL tool into Data Warehouse.

* Devised PL/SQL Stored Procedures, Functions, Triggers, Views and packages. Made use of Indexing,

Aggregation and Materialized views to optimize query performance.

* Developed logistic regression models (using R programming and Python) to predict subscription

response rate based on customer’s variables like past transactions, response to prior mailings,

promotions, demographics, interests and hobbies, etc.

* Created Tableau dashboards/reports for data visualization, Reporting and Analysis and presented it

to Business.

* Implemented FTP operations using Talend Studio to transfer files in between network folders as

well as to FTP server using components like tFileCopy, TFileAcrchive, tFileDelete,

tCreateTemporaryFile, tFTPDelete, tFTPCopy, tFTPRename, tFTPut, tFTPGet etc.

* Design and develop spark job with Scala to implement end to end data pipeline for batch processing
* Created Data Connections, Published on Tableau Server for usage with Operational or Monitoring

Dashboards.

* Knowledge in Tableau Administration Tool for Configuration, adding users, managing licenses and

data connections, scheduling tasks, embedding views by integrating with other platforms.

* Worked with senior management to plan, define and clarify dashboard goals, objectives and

requirement.

* Responsible for daily communications to management and internal organizations regarding status

of all assigned projects and tasks.

**Environment:**Hadoop Ecosystem (HDFS), Talend, SQL, Rampage, Tableau, Hive, Sqoop, Kafka, Impala, Spark, Unix Shell Scripting.