**Senior Data Engineer**

**Name: Mobin Shaik**

**Emai:** [shaikmobin453@gmail.com](mailto:shaikmobin453@gmail.com)

**Phone: 469 694 7746**

**LinkedIn:** [**linkedin.com/in/mobinsk**](https://www.linkedin.com/in/mobinsk)

**PROFESSIONAL SUMMARY**

* Data Engineering professional with solid foundational skills and proven tracks of implementation in a variety of data platforms. Self - motivated with a strong adherence to personal accountability in both individual and team scenarios.
* Over 9+ years of experience in Data Engineering, **Data Pipeline Design**, Development and Implementation as a **Sr. Data Engineer/Data Developer** and **Data Modeler**.
* Worked for 4 years with **AWS-BigData/Hadoop Ecosystem** in the implementation of **DataLake**.
* Hands on experience **Hadoop framework** and its ecosystem like Distributed file system **(HDFS)**, **MapReduce**, **Pig**, **Hive**, **Sqoop**, **Flume**, **Spark**.
* Strong experience in writing scripts using **Python API**, **PySpark API** and **Spark API** for analyzing data.
* Extensive experience in developing **Data warehouse** applications using **Hadoop**, **Informatica**, **Oracle**, **Teradata**, **MS SQL** server on **UNIX** and Windows platforms and experience in creating complex mappings using various transformations and developing strategies for **Extraction Transformation and Loading (ETL)** mechanism by using Informatica 9.x/8.x
* Extensively used Python Libraries **PySpark**, **Pytest**, **PyExcel**, **Boto3**, **Psycopg**, **embedPy**, **NumPy** and **Beautiful Soup**.
* Experience in **Google Cloud components**, **Google container builders** and **GCP** client libraries and cloud **SDK'S**
* Working experience with Functional programming languages like **Scala**, and **Java**.
* Extensive knowledge of **Data Modeling**, **Data Conversions**, **Data integration** and **Data Migration** with specialization in **Informatica Power Center**.
* Expertise in extraction, transformation and loading data from **heterogeneous** systems like **flat files**, **excel**, **Oracle**, **Teradata**, **MSSQL Server**.
* Extracting and modelling datasets from verity of **data sources** like **Hadoop** (using **Pig, Hive, Spark), Teradata** and **Snowflakes** for ad-hoc analysis and have fair understanding of AGILE methodology and practice.
* Developed and maintained mostly **Python** and some **Perl EL** scripts to scrape data from external web sites and load cleansed data into a **MySQL DB**.
* Enhancements to traditional data warehouse based on **STAR schema**, update **data models**, perform **Data Analytics** and Reporting using Tableau and extracted the data from **MySQL**, **Azure**, **AWS** into **HDFS** using Sqoop and created **Airflow** Scheduling scripts in Python.
* Design and Implement the Data Distribution Mechanism on SQL Server (**Transactional**, **Snapshot**, **Merge Replications**, **SSIS** and **DTS**).
* Have Extensive Experience in IT data analytics project, Hands on experience in migrating on premise ETLs to **Google Cloud Platform (GCP)** using cloud native tools such as **BIG query**, **Cloud Data Proc**, **Google Cloud Storage**, Composer.
* Very deep understanding and working knowledge on **Confluent Kafka** and **Apache Kafka** components.
* Experience working with **Cloudera**, **Amazon Web Services** (**AWS),** **Microsoft Azure** and **Hortonworks.**
* Hands on experience in **Normalization** and **De-Normalization** techniques upto 3NF for optimum performance in relational and dimensional database environments.
* Experienced in Amazon Web Services (**AWS**) and Microsoft **Azure**, such as **AWS EC2, S3, RD3**, **Azure HDInsight, Machine Learning Studio, Azure Storage, and Azure Data Lake.**
* Good understanding of Spark Architecture with **Data bricks**, **Structured Streaming**. Setting Up **AWS** and Microsoft **Azure** with **Data bricks** concepts like Event Hub and **Azure** **Data Bricks Data Lake** which provides fast and efficient processing of Big Data.
* Created **Hive External tables** and loaded the data into tables and query data using **HQL**.
* Interpret problems and provides solutions to business problems using **data analysis**, **data mining**, **optimization tools**, and **machine learning** **techniques** and **statistics**.
* Experience with **ETL** workflow Management tools like **Apache Airflow** and have significant experience in writing the **python** scripts to implement the workflow.
* Developed **Shell** and **Python** scripts to automate and provide Control flow to **Pig** scripts. Imported data from **Linux** file system to **HDFS.** Moreover, working closely with customers, cross-functional teams, research scientists, software developers, and business teams in an **Agile/Scrum** work environment to drive data model implementations and algorithms into practice.
* Strong experience in **Extraction, Transformation** and **Loading** (**ETL**) data from various sources into **Data Warehouse** and into **Azure Data Lake** using **Talend**, **SSIS**.
* Extensive experience in Text Analytics, developing different Statistical **Machine Learning**, **Data Mining solutions** to various business problems and generating data visualizations using **R, Python.**
* Experience in extracting source data from Sequential files, **XML** files, **CSV** files, transforming and loading it into the target **Data warehouse.**
* Knowledge of **Hadoop** administration activities using **Cloudera Manager** and **Apache Ambari**.
* Have good knowledge in Containers**, Docker,** and **Kubernetes** for the runtime environment for the **CI/CD** system to build, test, and deploy.
* Experienced in job workflow scheduling and monitoring tools like **Airflow, Oozie, TWS, Control**-**M** and **Zookeepe**r.
* Worked on Build Management tools like **SBT**, **Maven** and Version control tools like **Git.**

**TECHNICAL SKILLS**

* **Big Data Tools/** **Hadoop Ecosystem**: Map Reduce, Spark, Apache Airflow, Apache Flume, Nifi, HBase, Hive, Pig, Sqoop, Kafka, Oozie, Hadoop
* **Databases:** Oracle 12c/11g/10g, Teradata R15/R14, MySQL, SQL Server, MongoDB, Cassandra, DynamoDB, PostgreSQL, Teradata.
* **Data Modeling Tools:** Erwin 8/9, ER Studio v17
* **ETL/Data warehouse Tools:** Informatica and Tableau.
* **BI Tools:** SSIS, SSRS, SSAS, Power BI.
* **Programming Languages:** Python, PySpark, Scala, Java, C, C++, Shell script, Perl script, SQL
* **Cloud Platform:** Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP)
* **Containerization:** Docker, Kubernetes
* **Operating System:** Windows, Unix, Sun Solaris
* **Methodologies:** System Development Life Cycle (SDLC), Agile

**PROJECT EXPERIENCE**

**Client: Staples, Alpharetta, GA Oct 2021 – till now**

**Role: Senior Data Engineer**

**Responsibilities:**

* Designed and setup **Enterprise Data Lake** to provide support for various uses cases including Analytics, processing, storing, and reporting of voluminous, rapidly changing data.
* Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation and ensuring Integrity in a relational environment by working closely with the **stakeholders & solution architect**.
* Involved in the development of real time streaming applications using **PySpark**, **ApacheFlink**, **Kafka**, **Hive** on distributed **Hadoop Cluster**.
* Designed and developed Security Framework to provide fine grained access to objects in **AWS S3** using **AWS Lambda**, **DynamoDB**.
* Performed end-to-end Architecture & implementation assessment of various **AWS** services like **Amazon EMR**, **Redshift**, **S3**.
* Implemented the machine learning algorithms using **Python** to predict the quantity a user might want to order for a specific item so we can automatically suggest using **kinesis firehose** and **S3 data lake**.
* Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and **Amazon DynamoDB**.
* Used **Spark SQL** for Scala & amp, Python interface that automatically converts **RDD** case classes to schema RDD.
* Import the data from different sources like **HDFS/HBase** into **Spark RID** and perform computations using **PySpark** to generate the output response.
* Creating Lambda functions with **Boto3** to deregister unused **AMIs** in all application regions to reduce the cost for **EC2 resources**.
* Importing & exporting database using SQL Server Integrations Services (**SSIS**) and Data Transformation Services (DTS Packages).
* Coded **Teradata BTEQ** scripts to load, transform data, fix defects like SCD 2 date chaining, cleaning up duplicates.
* Developed reusable framework to be leveraged for future migrations that automates **ETL** from **RDBMS** systems to the **Data Lake** utilizing **Spark Data Sources** and **Hive** data objects.
* Conducted Data blending, Data preparation using Alteryx and **SQL** for Tableau consumption and publishing data sources to **Tableau** server.
* Developed **Kibana Dashboards** based on the Log stash data and Integrated different source and target systems into Elasticsearch for near real time log analysis of monitoring End to End transactions.
* Implemented **AWS** Step Functions to automate and orchestrate the **Amazon** **SageMaker** related tasks such as publishing data to S3, training **ML** model and deploying it for prediction.
* Integrated Apache Airflow with AS to monitor multi-stage **ML** workflows with the tasks running on Amazon **SageMaker**.

**Environment:** AWS EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Amazon SageMaker, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Snowflake, Apache Pig, Python, SSRS, Tableau.

**Client: Lowes, Rochester, MN Mar 2020 – Sep 2021**

**Role: Data Engineer**

**Responsibilities:**

* Responsible to create an orchestration layer across the core business areas On **Azure Analytics Services**, design and develop scalable, reliable, and performant **big data pipelines** to consume, integrate and analyse various datasets.
* Developed **PySpark** based **Spark** applications for performing data cleaning, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
* Developed Generic and Automated **python codes** for a set of **Data quality** Rules and Evaluation (data type casting issues)
* Implemented a Generic **Framework** for scheduling Daily Weekly/Monthly data loading from Curated to Consumption layer (Synapse)
* Worked on troubleshooting **spark** application to make them more error tolerant.
* Worked on fine-tuning **spark** applications to improve the overall processing time for the pipelines.
* Experienced in handling large datasets using Spark in Memory capabilities, using broadcasts variables in Spark, effective & efficient **Joins**, transformations, and other capabilities.
* Good experience with continuous Integration using **GIT**.
* Develop, construct, test end to end incremental architecture.
* Worked extensively with ADF for importing data from **Teradata** to **Azure Data Lake**.
* Experienced in **Databricks** platform where it follows best practices for securing network access to cloud applications.
* Using **Azure Databricks**, created Spark clusters and configured high concurrency clusters to speed up the preparation of high-quality data.
* Used Azure Databricks for fast, easy, and collaborative spark-based platform on **Azure**.
* Involved in creating delta tables, loading, and analysing data using **python scripts**.
* Implemented Partitioning for **Delta tables**. Analysed the SQL scripts and designed it by using **PySpark SQL** for faster performance.
* Designed, documented operational problems by following standards and procedures using **JIRA**.
* Experience in Developing Spark applications using Spark – SQL in **Databricks** for data extraction.
* Developed and operationalized batch and real-time models in **Sagemaker** with data science team.
* Used Reporting tools like **Power BI** to connect with **AAS** using Visual studio for generating daily reports of data.
* Collaborated with the infrastructure, network, database, application, and BA teams to ensure data quality and availability.

**Environment:** Spark, Python, PySpark, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse, Teradata, Snowflake, JIRA, Azure, Shell Scripting.

**Client: HCL/General Motors,** **Detroit, MI Oct 2018 – Jan 2020**

**Role: Big Data Engineer**

**Responsibilities:**

* Experience in building and architecting multiple **Data pipelines**, end to end **ETL** and ELT process for **Data ingestion** and transformation.
* Perform **Informatica Cloud Services**, **Informatica Power Center** Administration EL strategies and ETL Informatica mapping. Setting up of Secure Agent and connect different applications and its Data Connectors for processing the different kinds of data including unstructured (logs, click streams, Shares, likes, topics etc..), semi structured (XML, JSON) and structured like **RDBMS**.
* Worked extensively with **AWS** services like **EC2**, **S3**, **VPC**, **ELB**, Auto Scaling Groups, Route 53, **IAM**, **CloudTrail**, **CloudWatch**, **CloudFormation**, **CloudFront**, **SNS**, and **RDS**.
* Developed Python scripts to parse XML, Json files and load the data in **AWS Snowflake** Data warehouse.
* Outguessed the data from **HDFS** to **Azure SQL** data warehouse by building **ETL pipelines** using S worked on various methods including **data fusion**, **machine learning** and improved the accuracy of distinguished the right rules from potential rules.
* Design and Develop ETL Processes in **AWS Glue** to migrate Campaign data from external sources like S3, Parquet/Text Files into **AWS Redshift**.
* 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.
* Strong background in Data Warehousing, Business Intelligence and ETL process (**Informatica**, **AWS Glue**) and expertise on working on Large data sets and analysis
* Building a **Scala** and spark based configurable **framework** to connect common Data sources like **MYSQL**, Oracle, **Postgres**, SQL Server, Salesforce, Big query and load it in Big query.
* Applied required transformation using AWS Glue and loaded data back to **Redshift** and **S3**.
* Extensively worked on making **REST API** (application program interface) calls to get the data as JSON response and parse it.
* Experience in analyzing and writing SQL queries to extract the data in son format through Rest API calls with **API Keys**, **ADMIN Keys** and **Query Keys** and load the data into Data warehouse.
* Extensively worked on **Informatica** tools like source analyzer, mapping designer, workflow manager, **workflow monitor**, Mapplets, Worklets and repository manager.
* Building data pipeline ETLs for data movement to S3, then to Redshift. Designed and implemented ETL pipelines between from various Relational Data Bases to the Data Warehouse using **Apache Airflow**.
* Worked on **Data Extraction**, aggregations, and consolidation of **Adobe data** within AWS Glue using **PySpark**.
* Developed **SIS** packages to Extract, Transform and Load EL data into the **SQL Server database** from the legacy mainframe data sources.
* Worked on Building data pipelines in airflow in **GCP** for **ETL** related jobs using different **airflow operators**.
* Worked on Postman using **HTTP** requests to GET the data from **RESTful API** and validate the **API calls**.
* Hands-on experience with **Informatica Power Center** and power exchange in integrating with different applications and relational databases.
* Prepared dashboards using **Tableau** for summarizing Configuration, Quotes, Orders, and other e-commerce data.
* Created Informatica workflows and **IDQ mappings** for - Batch and Real Time.
* Created custom **T-SQL** procedures to read data from flat files to dump to SQL Server database using SQL Server import and export data wizard.
* Developed ETL python scripts for ingestion pipelines which run on **AWS** infrastructure setup of **EMR**, **S3**, **Redshift** and **Lambda**.
* Monitoring big query, **Dataproc** and cloud Data flow jobs via Stack driver for all environments.
* Configured EC2 instances and configured IAM users and roles and created S3 data pipe using Boto API to load data from internal data sources.
* Hands on experience with Alteryx software for ETL, data preparation for **EDA** and performing spatial and predictive analytics.
* Provided Best Practice document for **Docker**, **Jenkins**, **Puppet** and **GIT**.
* Expertise in implementing DevOps culture through **C/CD** tools like Repos, Code Deploy, Code Pipeline, **GitHub**.
* Developed various **Shell Scripts** for scheduling various data cleansing scripts and loading process and maintained the batch processes using **Unix Shell Scripts**.
* Backing up **AWS Postgres** to **S3** on daily job run on **EMR** using **Data Frames**. Developed server-based web traffic using **RESTful API's** statistical analysis tool using **Flask**, **Pandas**.
* Analyse various type of raw file like **Json**, **Cs**, **Xl** with **Python** using **Pandas**, **NumPy** etc.

**Environment:** Informatica Power Center 10./9., IDQ, AWS Redshift, Snowflake, S3, Postgres, Google Cloud Platform(GCP), MS SQL Server, Big query, Salesforce SQL, Python, Postman, Tableau, Unix Shell Scripting, EMR, GitHub.

**Client: HCL/ General Motors, Zapopan, Mexico Nov 2016 – Sep 2018**

**Role: ETL Developer/Data warehouse Engineer**

**Responsibilities:**

* Understanding the data warehouse and data mart models as well as being familiar with the business information contained in each with Active participation in **Data Warehouse** architecture and project discussions.
* Involved in Sprint planning, Daily Stand up calls, Retrospective and Review Meeting using Version one
* Develop and implement oozie workflows for Bigdata technologies and schedule them via crontab while monitoring their performance using Hue which is an opensource SQL workbench for DWH.
* Import data using **Sqoop** into **Hive** and **Hbase** from legacy source system (ADMIN SERVER).
* Develop complex mappings in **Informatica** Designer tool using transformations like lookup, aggregator, expression, router.
* Creates and schedules workflows for **mappings** in Informatica Workflow manager. These mappings and workflows are used for receiving data files from Admin server on a daily batch basis.
* Loading the admin files to **Valuations Annuities Greenplum DB** by performing a Change Data Capture and loading delta records in downstream Warehouse and mart.
* Optimized **Query Performance**, **Session Performance** and Reliability.
* Work with Data Governance Manager and **Data Stewards** to convert business requirements to Informatica **AXON/EDC** configuration.
* Create complex Structured Query Language (SQL) procedures and functions in **PG Admin** tool for Greenplum database.
* Created the Deployment groups and deployed the objects from one environment to the other using **Servicenow**.
* Worked with Informatica Best Practices guidelines to minimize performance issues.
* Created **AWS S3 buckets**, performed folder management in each bucket, managed cloud trail logs and objects within each bucket.
* Completed POC on usability of **AWS** using **EC2** and **S3** storage and lambda functions.
* Developed, Code Reviewed, fixed bugs, tested and deployed SSIS packages using **SQL Server 2012/2008 R2** Business Intelligence Development Studio.
* Transferred data from various data sources/business systems including **DB2**, MS Excel, MS Access, Flat Files etc. to SQL Server using **SSIS** Packages and using various features like Excel source, Flat file source, transformation etc. Created derived columns from the present columns for the given requirements.

**Environment:** Informatica Power center 10.1.1, Oracle 12C, PLSQL Developer, Greenplum 9.6, PG Admin, YARN, MapReduce, Pig, Hive, HBase, Sqoop, Oozie, Hue, Linux, Informatica AXON, EDC, MSBI 2012/2008

**Client: HCL, Hyderabad, India March 2014 – Nov 2016**

**Role: SQL/ BI Developer**

**Responsibilities:**

* Designed, developed, and maintain BI solutions using SQL, including **data warehouse** and **data mart** structures.
* Collaborated and coordinated with business analysts to gather requirements for implementing data mart.
* Created required SQL procedures, functions, triggers, and views to supports **OLTP** and **ETL** application.
* Worked with various SSIS control flow tasks and data transformations like Data conversion, Derived, Column, Look-up, Fuzzy Look-up, Conditional Split, Aggregate transformations etc. as part of ETL.
* Created Event Handlers, Package Configurations, Logging, System and User-defined Variables Incorporated control for solution source control, project coordination, and task management.
* Created AD-Hoc reports using Report Builder and maintained Report Manager for **SSRS**.
* Built effective queries for high-performance reporting and rendered them to HTML, XML, PDF and Excel format using **SSRS**.
* BI team is responsible for development, enhancement & maintenance of various internal and external reports/applications to support claims, members, providers, compliance related processes.
* Migrated data from on-premises data servers onto **Microsoft Azure cloud platform** utilizing **Azure data factory** and experience with Data Lake.
* Developing SSRS reports on SQL Server and have very good understanding of **SSAS**, **OLAP** cube and Architecture.
* Used **ETL** (**SSIS**) to develop jobs for extracting, cleaning, transforming, and loading data into data warehouse.
* Designed and developed various analytical Dashboards from multiple data sources by blending data on a single worksheet in **Tableau Desktop**.
* Created **Tableau dashboards** with Hierarchies, interactive views, drill downs with user level security.
* Worked in Tableau environment to create dashboards like weekly, monthly, daily reports using tableau desktop and published them to server.
* Designed a database optimization plan like creation of table partitions and the use of appropriate Indexes.

**Environment:** Microsoft SQL Server 2012, SSRS, SQL Server Integration Services (SSIS) 2012, Microsoft SQL Server Management Studio, BIDS, T-SQL, SAP, C#, Azure Data Factory, Tableau Desktop V8/9, TFS, Azure