**Syed Sohel**

**Chicago, IL**

**Email:** [**sohelpasha637@gmail.com**](mailto:sohelpasha637@gmail.com)

**LinkedIn:** [linkedin.com/in/sohel-syed-5721aa277](https://www.linkedin.com/in/sohel-syed-5721aa277)

**Professional Summary:**

Data Engineer with 6 years of experience specializing in designing, developing, and maintaining data solutions. Expertise in cloud-based environments, big data technologies, and software development. Proven track record of delivering high-quality products and services in fast-paced, data-driven organizations. Skilled in collaborating with cross-functional teams to drive innovation and solve complex problems with a strong background in lending industries, specifically home lending. Hands-on experience in migrating data from Hadoop to Google Cloud Platform (GCP) and BigQuery.

* Extensive experience in designing both Front end & Backend applications using Java, J2EE Web frameworks, JSP, JSTL, HTML, CSS, Angular 2 and Angular4, AJAX, JavaScript, jQuery, Bootstrap.
* Extensive experience in various Java/J2EE technologies including Core Java, J2EE, Servlets, JSP, JDBC, Struts, spring and hibernate.
* Extensive Experience working with spring, Struts, JSF and O/R mapping Hibernate frameworks. Familiarity with Home Lending and Consumer Lending industry.
* Advance knowledge of NoSQL, SQL Queries using Hibernate framework with Spring ORM in interaction with the Relational Database Management System (RDBMS).
* Worked with Apache Solr to implement indexing and wrote Custom SOLR query segments to optimize the search.
* Developed a knowledge object using Python and R to calculate the life expectancy of pancreatic cancer patients.
* Proficient in ETL development using Talend, ensuring accurate and efficient data migration. Developed and implemented ETL code using Talend, ensuring adherence to specifications and design guidelines.
* Processed the Web server logs by developing multi-hop flume agents by using Avro Sink, loaded into MongoDB for further analysis, extracted files from MongoDB through Flume, and processed.
* Used OOZIE Operational Services for batch processing and scheduling workflows dynamically.
* Used Zookeeper to coordinate the servers in clusters and to maintain the data consistency.
* Used Spark and Spark-SQL to read the parquet data and create the tables in hive using the Scala API.
* Proficient in writing Terraform configuration files (.tf) to define infrastructure components, including compute instances, storage buckets, networking, load balancers, and other Google Cloud services.
* Good experience in CI/CD pipeline management through Jenkins. Automation of manual tasks using Shell scripting.
* Seeking to leverage my expertise in extracting, transforming, and loading marketing and customer data to contribute to the success of Adobe CDP as a Senior Data Engineer.
* Hands on databases like Oracle, MS SQL, DB2 and developing in RDBMS that includes SQL queries, Stored procedures, and triggers.
* Strong knowledge in working with UNIX/LINUX environments, writing shell scripts and PL/SQL Stored Procedures.
* Experienced in using source code, control systems such as GIT, SVN, and CVS. Diverse experience utilizing tools in N-tier and Microservices architecture applications using Spring Boot, My SQL, Restful Web Services.
* Implemented the use of Amazon EMR for Big Data processing among a Hadoop Cluster of virtual servers on Amazon related EC2 and S3.

**WORK EXPERIENCE**

**Blue Cross Blue Shield (Chicago, IL)**

**Hadoop/Data Engineer January 2022 - Present**

* Designed and engineered complex data solutions, ensuring consistency across multiple flows and systems. Developed processes for data transformation, data structures, metadata, data quality controls, dependency management, and workload management.
* Defined internal controls and identified gaps in data management standards adherence, working with partners to develop plans for improvement.
* Skilled in collaborating with business teams, performing data modeling, and leveraging SQL expertise to unlock valuable insights from complex datasets.

**Key Contributions:**

* Successfully implemented Hadoop stack and storage technologies, including HDFS, MapReduce, Yarn, HIVE, Sqoop, Impala, Spark, Flume, Kafka, and Oozie.
* Utilized Big Data Enterprise architecture (Cloudera preferred) to design and optimize data solutions. Applied strong analytical capabilities and algorithmic expertise to drive effective data analysis.
* Migrate data from Hadoop to GCP BigQuery, ensuring data integrity and optimizing performance.
* Leveraged HBase, RDBMS, SQL, ETL, and data analysis techniques to handle diverse data requirements. Utilized NoSQL technologies such as Cassandra and MongoDB for efficient data processing.
* Demonstrated ability to work with Google Cloud APIs and services to integrate them into Terraform deployments.
* Developed Unix/Linux scripts for automation and employed scheduling tools like Autosys.
* Design and develop scalable data pipelines using GCP services such as BigQuery, GCS, Tekton, and Airflow.
* Collaborated with cross-functional teams, including Architects, Developers, Business/Data Analysts, QA, and client stakeholders, to ensure successful project delivery.
* Worked as a Spark Expert and performance Optimizer. Written MapReduce (Hadoop) programs to convert text files into AVRO and loading into Hive (Hadoop) table.
* Extensive eDesigned and developed scalable ETL pipelines using Hive and MySQL/XDB to process large volumes of data for real-time analytics.
* Leveraged S3, Apache Parquet, and Apache Iceberg to store and process large volumes of data efficiently.
* Proficient in performing ETL operations using GCP tools such as BigQuery, Dataflow, Dataproc, and Airflow.
* Utilize SQL expertise to create and optimize complex queries for data extraction, transformation, and analysis
* Collaborated with the RL and partner teams to understand their objectives and developed data solutions to support their initiatives. Experience on building REST API using Java and Python.
* Developed and optimized Spark applications with PySpark to process large-scale data sets. Orchestrated workflows on AWS using tools like AWS Step Functions and AWS Data Pipeline.
* Designed and optimized data warehouse solutions using Redshift, Snowflake, and ETL tools, ensuring fast and accurate reporting.
* Proficiency in Excel, including modeling in Excel and utilizing Power Pivot for advanced data analysis and reporting.
* Experienced in utilizing GCP services such as BigQuery, GCS, Tekton, and Airflow for building scalable and reliable data pipelines.
* Traced logic and application/query flow through logs using Spark/Impala
* Administered Cloudera Manager for platform monitoring and management.
* Implemented scripting/programming solutions using Shell/Python/JavaDesigned and implemented data warehousing solutions for efficient data storage and retrieval.
* Leveraged Lambda functions to build serverless data processing workflows and automate tasks. Managed and optimized AWS RDS databases for high performance and scalability.
* Worked with Kubernetes and containerization technologies to deploy and manage data applications. Worked with AWS Workflow Orchestration tools to schedule and manage data workflows.
* Developed and implemented ETL code using Talend, ensuring adherence to specifications and design guidelines.
* Develop and maintain data pipelines for efficient data ingestion, transformation, and storage.
* Implement data quality and data governance processes to ensure data accuracy and consistency.
* Developed and maintained ETL workflows using Talend, ensuring efficient data extraction, transformation, and loading.
* Lead the design, development, and implementation of scalable data and analytics solutions, ensuring high performance, reliability, and maintainability.
* Collaborate with cross-functional teams to understand business requirements and translate them into technical solutions.
* Worked in an agile environment, participating in scrum meetings and delivering high-quality results within sprint timelines.
* Familiarity with agile methodologies and working in an agile development environment.
* Collaborated with stakeholders to gather requirements and provide technical expertise in ETL processes.
* Support code/design analysis, strategy development and project planning. Create reports for the BI team using Sqoop to export data into HDFS and Hive.
* Implemented Alteryx macros and automation to optimize and standardize data processing tasks, increasing efficiency and productivity.
* Demonstrated expertise in writing complex SQL queries to identify and measure non-genuine software base, applying business logic for categorization and identification of opportunities.
* Involve in Requirement Analysis, Design, and Development. Export and Import data into HDFS, HBase and Hive using Sqoop.
* Extracted and updated the data into HDFS using Sqoop import and export.
* Developed HIVE UDFs to incorporate external business logic into Hive script and developed join data set scripts using HIVE join operations.
* Involved in creating HDInsight cluster in Microsoft Azure Portal also created Eventshub and Azure SQL Databases.
* Worked on a clustered Hadoop for Windows Azure using HDInsight and Hortonworks Data Platform for Windows.

**Environment:**

Mongo, Adobe, Cloudera CDP 7.1.X, HDFS, MapReduce, Hive, Talend, Pig, ETL, Flume, Spark, Scala, Elastic search, Shell scripting, UNIX. NoSQL Databases: MongoDB, Cassandra, Redshift, Snowflake, Linux

**J.P. Morgan Chase (Hyderabad, India)**

**Data Engineer May 2017 - Dec 2021**

* Designed and built data applications using relational and NoSQL databases, open data formats, and programming languages such as Python, Scala, and others.
* Developed data pipelines (ETL and ELT) with batch and streaming ingestion, error handling, loading, and data transformation. Involved in Data Modelling, engineering, and Design of traditional RDBMS and Cloud systems.
* Proficient in performing ETL operations using GCP tools such as BigQuery, Dataflow, Dataproc, and Airflow.
* Worked with big data technologies such as Spark, Hadoop, and MapReduce.
* Implemented analytics solutions to extract valuable insights from data.
* Ensured data quality, reliability, availability, and scalability in existing and new applications. Proficient in ETL development using Talend, ensuring accurate and efficient data migration. Developed and implemented ETL code using Talend, ensuring adherence to specifications and design guidelines. .
* Applied DevOps concepts, cloud architecture, and Azure DevOps Operational Framework.
* Designed solutions for data quality, observability, metadata management, data lineage, and data discovery. Built micro-services-oriented architecture and REST APIs.
* Leveraged open-source frameworks and employed continuous delivery and infrastructure as code practices. Experience working on Data Warehouse and Data Modelling to Design and architect the process of data flow.
* Utilized monitoring portals like Splunk and Application Insights for operational insights.
* Demonstrated understanding of security protocols and products, including Active Directory, SAML, and OAuth.
* Proficient in Azure Network and operational portals.
* Create visualizations and reports for the business intelligence team, using Tableau and Ingested data from RDBMS and performed data transformations, and then export the transformed data to MongoDB as per the business requirement.
* Wrote check code in Groovy to verify the correctness of map data. Analyzed failures found by the checks to determine exactly what was wrong. Assisted other developers with learning Groovy and advised them on writing efficient code.
* AWS Migrated Hive QL queries on structured data into Spark QL to improve performance.
* Involved in improving the performance and optimization of the existing algorithms using Spark. Used SQL, Hive for querying the Hive and ORACLE Databases.
* Developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Handled large datasets using Partitions, Spark in-memory capabilities, broadcasts in Spark, effective & efficient Joins, transformations.
* Well experienced in handling Data-Skewness in Spark-SQL using Cloudera CDP 7.1.X.
* Implemented MLLIB in Spark and used K-means algorithm to cluster the data available in hive tables.
* Worked with Apache Solr to implement indexing and wrote Custom SOLR query segments to optimize the search.
* Used OOZIE Operational Services for batch processing and scheduling workflows dynamically.
* Used Zookeeper to coordinate the servers in clusters and to maintain the data consistency.
* Developed UNIX shell scripts to load large number of files into HDFS from local file system.

**Environment:**

Hadoop, Hive, Data Modelling, SQL, Map Reduce, Sqoop, Spark, Cloudera CDP 7.1.X, Eclipse, Maven, Java, Python, Jira, REST API, Home lending, Maven, JSP, Spring, Hibernate, XML, SQL, JUnit, Log4J, MySQL, Oracle, Data Warehouse and DB2.

**Skills:**

Data software development, big data technologies: 5+ years

Designing and building data applications: 4+ years

Cloud DevOps concepts, Cloud Services and Architecture: 2+ years (Azure/AWS/GCP)

Open-source data tools and frameworks or .NET Core, asp.Net, Angular, or Express: 1+ years

Relational & NoSQL databases, open data formats, Python, Scala, and other frameworks: Experienced

Building data pipelines with batch or streaming ingestion, error handling, loading, and transforming data: Experienced

Data Warehouse, Data Modelling, Spark, Hadoop, and MapReduce: Experienced

ETL - TALEND: Experienced

Python, PySpark, Spark, Scala: Experienced

DevOps Concepts, Cloud Architecture, Azure DevOps Operational Framework, Pipelines, Kubernetes: Advanced

Data quality, observability, metadata management, data lineage, and data discovery: Advanced

Micro-services-oriented architecture, extensible REST APIs: Advanced

Monitoring Portals: Splunk, Application Insights: Experienced

Security Protocols & Products: Active Directory, Windows Authentication, SAML, OAuth: Advanced

Azure Network and operational portals: Advanced

**Education**

**MS in Business Analytics** from Trine University (2023)

**B. TECH** from Osmania University – Computer Science. (2014)