**Joshi Gandham**

**Mobile: +1 936-386-2220**

**Joshipg09@gmail.com**

**PROFESSIONAL SUMMARY:**

* **Overall 8 years’** experience in IT Industry as **Data Engineer** with a wide range knowledge of big data analysis, data cleaning/visualizing/exploring, building ETL pipelines, and cloud computing.
* Strong experience on Apache Hadoop ecosystem components like **MapReduce**, **HDFS**, **HBase**, **Hive**, **Sqoop**, **Pig**, **Oozie**, **Zookeeper**, **Flume**, **Spark**, **Scala,** and **Python.**
* Extensive experience in designing, **building**, and maintaining **data** **pipelines** on **AWS** using services such as **S3**, **Glue**, and **EMR**.
* Proficient in optimizing data storage and retrieval using **AWS** services such as **Redshift**, **Lambda**, **DynamoDB**, and **Aurora**.
* Experience designing and implementing large-scale data warehousing solutions using **GCP** tools like Google **BigQuery** and Google **Cloud** **Storage**.
* Proficiency in batch and stream processing using GCP tools like Google **Cloud** **Dataflow** and Google **Cloud Pub/Sub**.
* Experience working with Hadoop and Spark on GCP using services like Google **Cloud** **Dataproc** and Google **Cloud** **Composer**.
* Have good experience designing cloud-based solutions in Azure by creating **Azure** **SQL** **database**, setting up Elastic pool jobs and design tabular models in Azure analysis services.
* Have extensive experience in creating pipeline jobs, schedule triggers using **Azure** **data** **factory**.
* Strong experience in migrating other databases to **Snowflake**.
* Experience with **Snowflake** cloud data warehouse and **AWS S3** bucket for integrating data from multiple source system which include loading nested **JSON** formatted data into snowflake table.
* Created and maintained **CI/CD** (continuous integration & deployment) pipelines and apply automation to environments and applications. Worked on various automation tools like **GIT** and **Terraform**.
* Have Data warehousing experience in Business Intelligence Technologies and Database with Extensive Knowledge in **Data analysis, TSQL queries, ETL & ELT Process, Reporting Services (using SSRS, Power BI) and Analysis Services using SQL Server.**
* Have good understanding in best practices of **Enterprise Data Warehouse** and involved in Full life cycle development of **Data Warehousing.**
* Expertise in Extraction, Transformation, loading data from **Oracle, DB2, SQL Server, MS Access, Excel, Flat Files and XML using Informatica and Talend**
* Experience in application of various data sources like **Oracle** **SE2**, **SQL** **Server**, **Flat** **Files** and Unstructured files into a **Data** **warehouse**.
* Proficient in using **Boto3** to interact with diverse AWS services, including S3, EC2, Lambda, SQS, and DynamoDB.
* Strong **SQL** development skills including writing **Stored Procedures, Triggers, Views**, and **User Defined functions.**
* Experienced with containerization and orchestration services like **Docker, Kubernetes**.
* Expertise in Build Automation and Continuous Integration tools such as **Maven, Jenkins**.
* Worked on Scala code base related to Apache Spark performing the Actions, Transformations on **RDDs**, Data Frames & Datasets using **Spark SQL** and Spark Streaming Contexts.
* High proficiency in requirement gathering and data modeling including design and support of various applications in Online Transactional processing (**OLTP**) and Online Analytical Processing (**OLAP**), Data Warehousing and ETL Environment.
* Prepared **ETL mapping** Documents for every mapping and **Data Migration** document for smooth transfer of project from development to testing environment and then to production environment.
* Extract, transform and load the data from different formats like **JSON**, a **Database**, and expose it for ad-hoc/interactive queries using **Spark** **SQL**.
* Experience of using **Jira** to capture/track/resolve issues, custom workflows, and projects configuration.
* Well versed with **Agile with SCRUM, Waterfall Model** methodologies.
* Excellent communication and written skills with A self-motivated learner and adequate with challenging projects and work in ambiguity to solve complex problems independently or in the collaborative team.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming languages** | Python, Java, SQL |
| **Big Data technologies** | Hadoop, Spark, EMR, Glue, Athena, Databricks, HDInsight, Data Lake Analytics, Pig, Sqoop, Flume, Oozie, ZooKeeper |
| **AWS Services** | Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Cloud Watch, Amazon EBS, Amazon CloudFormation, Athena |
| **Google Cloud Platform (GCP)** | Google BigQuery , Google Cloud Storage, Cloud Dataflow, Cloud Pub/Sub, Dataproc, DataFusion, Cloud Composer, Cloud SQL, Cloud Datastore and Google Cloud Memorystore |
| **Microsoft Azure** | Azure Data Factory, Azure Data Lake, Databricks, Active Directory, Application Insights, Azure SQL DB, SQL data warehouse, Synapse Analytics (DW) |
| **Hadoop Ecosystem** | HDFS, Map Reduce, YARN, Hive, Zookeeper, Sqoop, Oozie, Apache Cassandra, Apache Airflow, Flume, Spark |
| **Data modeling** | ER Modeling, Dimensional Modeling, Normalization |
| **Database management** | MySQL, PostgreSQL, MongoDB, Cassandra |
| **ETL** | Informatica, Data Stage, SSIS |
| **Tools** | Git, Jenkins, JIRA, Docker, Kubernetes |

**PROFESSIONAL EXPERIENCE:**

**Client: Oncor, Dallas, TX. Jun 2022 – Present AWS Data Engineer:**

**Responsibilities:**

* Designed and developed **ETL** (Extract, Transform, Load) pipelines using AWS services such as **Glue**, **Lambda**, **and** **S3** to extract data from various sources, transform it and load it into a data warehouse.
* Build and maintain scalable and highly available data solutions using AWS services such as **Redshift**, **RDS**, and **DynamoDB**.
* Extensively worked with the application data present in AWS **S3** storage for performing required transformations.
* Configured and deployed production-ready data pipelines using AWS Glue, S3 and Lambda to maximize data extraction, transformation, and loading.
* Designed and implemented data ingestion pipelines using Apache **Airflow** to efficiently transfer and process data from various sources.
* Created efficient **DAGs** to represent data workflows, defining task dependencies and execution order.
* Integrated Apache **Airflow** with Amazon S3 to store and retrieve data, ensuring seamless data flow between different stages of the pipeline.
* Optimized **DAGs** for parallel execution, improving data processing performance and reducing processing times.
* Performed transformations on the data using **Glue** **jobs** with the **pyspark** script and stored transformed data to **S3**.
* Integrated Apache **Airflow** with AWS Glue Data Catalog for centralized metadata management and cataloging of data assets within the data pipeline.
* Optimized **EMR** usage and established scalable, fault-tolerant **ETL** **pipelines** to facilitate data extraction from multiple sources for storage in the **S3** data lake, increasing output efficiency.
* Set up CI/CD pipelines to automate testing and deployment of **Airflow DAGs**, enhancing the development and release process.
* Implementing **AWS Big Data services** such as **EMR, Kinesis**, and **Athena** to process large volumes of data.
* Utilized **AWS** **Athena** interactive querying service for running **Ad**-**hoc** queries to perform **data** **analysis** on transformed data
* Executed **Python** scripts to automate ETL jobs and performing a required transformation on the large amount of data using **AWS EMR**
* Performed appending of weekly data present in AWS **S3** using **pyspark** applications with **AWS EMR** (**Elastic MapReduce**)
* Worked on AWS **Redshift** data warehouse service for running **SQL** queries on the large amount of transformed data present in **S3**.
* Employed **Boto3** to automate the creation and management of **EC2** instances.
* Utilized **AWS** **Glue** **Data Catalog** for data discovery and metadata management.
* Developed a **Boto3** script to dynamically scale **AWS** **Lambda** functions based on workload.
* Created and maintained data tables, databases, and schemas in the **AWS** **Glue Data Catalog**.
* Developed and maintained AWS **Lambda** functions using **Python** for real-time data processing.
* Worked with other members of the engineering team for building **data** **pipelines** and to automate the processing of a large amount of data using AWS **cloud** **watch** and **Lambda** **services.**
* For Data Governance worked on AWS Identity and Access Management (IAM) for fine-grained access control to data resources.
* Configured and managed **AWS SNS** topics to facilitate seamless communication between microservices and distributed systems.
* Integrated **SNS** with other AWS services such as **Lambda**, **EC2**, and **S3** to trigger automated workflows and actions.
* Integrated **SQS** with AWS **Lambda** to process messages asynchronously, enabling serverless processing.
* Utilized **SQS** delay queues to schedule the delivery of messages, optimizing system workflows.
* Worked with AWS **CloudFormation** templates to automate the deployment and management of **SNS** and **SQS** resources.
* Conducted performance tuning and optimization of **Scala**-based data processing jobs.
* Used **AWS Cloud watch**, **Lambda** services for automatic triggering of the events and invoking the underlying code to execute
* Integrated AWS **ECS** with AWS **CloudWatch** for monitoring and continuous deployment.
* Designed and maintained data profiling and data cleansing processes to identify and rectify data anomalies and errors.
* Created data validation scripts and automated checks to ensure data integrity and compliance with defined data quality standards.
* Adhering to best practices and industry standards for **data warehousing** and **data modeling**.
* Teamed up with data analysts by providing queried data from **AWS** **Athena** and **Redshift** and assisted in the creation of **visualizations** using **Tableau**. Worked closely with them to optimize the efficiency of data for **Tableau dashboards**.
* Used **Agile** practices such as pair programming, **code** **reviews**, and **continuous** **integration** to ensure code quality and maintainability.

**Client: Maximus, Princeton, NJ. Feb 2021 – May 2022 GCP Data Engineer:**

**Responsibilities:**

* Designed and implemented a data pipeline to process and store large volumes of data using **GCP** technologies such as **Apache** **Beam** and **Dataflow**.
* Developed an **ETL** process to transform raw data into a structured format using tools such as **Cloud** **Dataprep** and **Cloud** **Functions**.
* Utilized **Apache Beam** to construct complex data transformation, ensuring data quality and consistency. workflows
* Developed and deployed **Python**-based data pipelines on Google Cloud **Dataflow** for high-volume data processing.
* Built a data warehouse on **GCP** using **BigQuery** and optimized it for performance and cost.
* Developed a complex **SQL** query to aggregate and analyze transaction data, resulting in a 20% improvement in identifying patterns.
* Implemented scheduled exports and backups of **BigQuery** datasets to ensure data durability and recovery.
* Integrated new third-party data sources into the data pipeline using **APIs** and cloud storage services.
* Developed and scheduled data processing jobs using **Dataproc** and Apache **Spark**.
* Implemented **data** **validation** and cleaning procedures to ensure integrity and improved data quality by 20%.
* Developed automated **workflows** using **Cloud** **Composer** to reduce manual intervention and improve efficiency by 50%.
* Utilized **Dataproc** to process and analyze large-scale datasets.
* Utilized **Python** along with Google Cloud Storage (**GCS**) to store and retrieve data.
* Utilized **Dataproc** autoscaling to dynamically adjust cluster size based on workload demands.
* Developed data profiling and cleansing tools using GCP services such as **Cloud** **Dataprep** and **Data** **Loss** **Prevention** (**DLP**) to ensure data accuracy and compliance.
* Created and managed a Data Catalog using Google **Cloud Data Catalog**.
* For Data Catalog implemented data tagging and annotations for improved data classification.
* Reduced data processing latency by 40% by implementing GCP's autoscaling feature, which added or removed processing nodes as needed.
* Implemented **PySpark** jobs on GCP's Cloud **Dataflow** to perform batch data processing.
* Utilized **PySpark** on GCP's **Dataproc** service to process and analyze large-scale datasets in a distributed and scalable manner.
* Integrated **Dataproc** with **Git** repositories for version control.
* Worked with cross-functional teams to identify key performance indicators (**KPIs**) and established a data pipeline that provided real-time **KPI** **tracking** and **reporting**.
* Implemented data governance policies and controls using Cloud Identity and Access Management (IAM).
* Created data models and dashboards using Looker or **Data** **Studio** to help stakeholders understand and make decisions based on the data.
* Worked with PowerShell and **Linux** scripts for file transfer, emailing and other file related tasks.
* Monitor and troubleshoot data pipelines for performance, scalability, and reliability issues.
* Conducted data profiling and cleansing processes to improve the quality of critical datasets.
* Improved data security and access control by implementing role-based access controls and encryption.
* Documented the data pipeline architecture and processes for future reference.
* Implemented a **CICD** pipeline using **Git** and **BitBucket** for version control and code repository management, enabling continuous delivery of data solutions across multiple environments.
* Collaborated with data analysts and business stakeholders to understand their data requirements and provide them with data solutions.

**Client: Fox Corporation, San Francisco, CA. Feb 2020– Jan 2021 Azure Data Engineer:**

**Responsibilities:**

* Designed and Configured **Azure** **Cloud** relational servers and databases analyzing current and future business requirements.
* Designed and developed complex ETL mappings using **Informatica** PowerCenter to extract data from multiple sources (relational databases, flat files) and load it into target data warehouses.
* Utilized **Informatica** Data Quality (IDQ) to profile and cleanse data from various sources, identifying and resolving data quality issues
* Migrated **Snowflake** database to Windows **Azure** and updating the Connection Strings based on requirement.
* Developed **Talend MDM** jobs to populate the claims data to data warehouse - **star** **schema**, **snowflake** **schema**, **Hybrid** **Schema**.
* Experience in developing **Spark** applications using **Spark-SQL** in Databricks for data extraction, transformation, and aggregation from multiple file formats for **Analyzing**& **transforming** the data to uncover insights into the customer usage patterns.
* Developed **Data** **Factory** pipelines that triggered **Databricks** notebooks for data processing, transforming data from Azure **Blob** Storage and storing the results in Azure **SQL** Database.
* Have extensive experience in creating **pipeline** **jobs**, scheduling triggers, Mapping data flows using **Azure** **Data** **Factory**(**V2**) and using **Key** **Vaults** to store credentials.
* Good experience in creating **Elastic** **pool** databases and schedule **Elastic** **jobs** for executing **TSQL** procedures.
* Extract Transform and Load data from sources Systems to **Azure Data Storage** services using a combination of **Azure Data factory, T-SQL,** **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 **Azure Databricks.**
* Used **Python** with Azure **Databricks** to process large volumes of data efficiently.
* Hands on experience on developing SQL Scripts for automation
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the **Spark** **Databricks** cluster
* Developed and enforced data governance policies and procedures, ensuring data compliance with regulatory standards like GDPR, HIPAA.
* Have good experience in logging defects in **Jira** and **Azure Devops** tools.
* Worked on creating correlated and non-correlated sub-queries to resolve complex business queries involving multiple tables from different databases.
* Implemented Disaster Recovery and Failover servers in Cloud by replicating data across regions.
* Good understanding of Spark Architecture including **spark core**, **spark SQL**, **DataFram**e, **Spark streaming**, **Driver Node**, **Worker N**ode, **Stages**, **Executors** and **Tasks**, Deployment modes, the Execution hierarchy, fault tolerance, and collection.
* Developed custom data connectors and integration solutions in **Scala** to ingest data from various sources.
* Created branches, labels and performed merges in stash and **GIT**.
* Worked on **Linux** scripts for file transfer and emailing tasks from FTP/SFTP.
* Experience working in **Agile** development environments, including **Scrum** or a hybrid approach.
* Familiarity with **Agile** principles such as continuous delivery, incremental development, and cross-functional collaboration.
* Participated in daily stand-up meetings, sprint retrospectives, and backlog grooming sessions to track progress, resolve issues, and improve processes. Used **Jira** to track tasks and update status.

**Client: ION Broadband, Hyderabad, India Sep 2014 – Dec 2019 Hadoop Data Engineer:**

**Responsibilities:**

* Hands on experience in Spark and **Spark Streaming** creating **RDD** & applying operations, transformations, and Actions.
* Used **Python** with Apache **Spark** for distributed data processing on **Hadoop** clusters.
* Used Spark and Spark-SQL to read the **parquet** data and create the tables in hive using the **Scala**

**API**.

* Developed Spark code using Scala and **Spark-SQL** for faster processing and testing.
* Designed and implemented complex data processing pipelines using **Scala** and Apache **Spark.**
* Experienced Good understanding of NoSQL databases and hands on work experience in writing applications **No** **SQL** Databases **HBase**, **Cassandra** and **MongoDB**.
* Utilized **Control**-**M** to schedule and automate data pipelines, ensuring timely and reliable execution of data processing and ETL jobs.
* Worked on Hadoop cluster and data querying tools Hive to store and retrieve data.
* While developing applications involved in complete Software Development Life Cycle (**SDLC**).
* Involved in the design, development and testing phases of application using **AGILE** methodology.
* Reviewing and managing **Hadoop** log files by consolidating logs from multiple machines using **flume**.
* Performed various Parsing technique's using spark API'S to cleanse the data from **Kafka**.
* Developed **Oozie** workflow for scheduling **ETL** process and **Hive Scripts**.
* Responsible for designing and implementing ETL process using **Talend** to load data from Worked extensively with **Sqoop** for importing and exporting the data from **HDFS** to Relational Database systems/mainframe and vice-versa.
* Responsible for loading **Data pipelines** from web servers and **Teradata** using **Sqoop** with **Kafka** and **Spark Streaming API**.
* Developed **Kafka** producer and consumers, **Cassandra** clients and Spark along with components on **HDFS**, **Hive**.
* Implemented job dependencies, conditions, and triggers in **Control-M** to ensure proper sequencing and execution of data tasks.
* Populated **HDFS** and **HBase** with huge amounts of data using **Apache Kafka**
* Experience in data pipelines using **Kafka** and **Akka** for handling large terabytes of data.
* Written shell scripts that run multiple Hive jobs which helps to automate different Hive tables incrementally which are used to generate different reports using **Tableau** for the Business
* Developed automated **Tableau** reports and dashboards to streamline data analysis and reduce manual effort.
* Worked on **Linux** scripts for spliting group of files to small files and for file tansfering automation.
* Worked on the concept of **GIT** cloning to create repository.
* Manipulate, serialize, model data in multiple forms like JSON, XML. Involved in setting up **MapReduce 1** and **MapReduce 2.**
* Prepared **Avro** **schema** files for generating Hive tables and Created **Hive tables** and loaded the data into tables and query data using **HQL**.
* Installed and Configured **Hadoop** **cluster** using **Amazon Web Services** (AWS) for POC purpose

**Education:**

* Master’s in Computer and Information Sciences, **Gannon** **University**, Erie, PA
* Bachelor’s in Electronics and Communication Engineering, **Jawaharlal Nehru Technological University**, Hyderabad, India.

**Certifications:**

* SnowPro Core Certification

Credly URL: **https://www.credly.com/badges/2b1ab855-9933-4210-a5ab-ee320544eb3e/public\_url**

* AWS Certified Solutions Architect – Associate Certification

Credly URL: **https://www.credly.com/badges/192bb1f5-7f87-4257-a174-b71cf71e7a5d/public\_url**