

**Sai Gabbula**

**Data Engineer | Email:** [**saigabbula89@gmail.com**](mailto:saigabbula89@gmail.com) **| Contact: +1(336)252-8989**

|  |  |
| --- | --- |
| PROFESSIONAL EXCELLENCE SUMMARY: |  |

* 9+ years of experience working with almost all **Hadoop ecosystem components, AWS cloud services, Microsoft Azure, Google** **Cloud** **Platform, Apache Spark,** and 6+ years working experience and strong background in designing, developing, and deploying complex data integration solutions using Talend (Data Integration/Big Data Integration (6.1/5.x), Talend Data Quality) and Informatica (PowerCenter, IICS, IDQ, MDM).An enthusiastic and dedicated person who can accurately perform challenging tasks that demand precision and exceptional analytical abilities. Strengths include time management, problem-solving, and decision-making skills with the ability to set priorities and affirmatively produce results.
* Experience with Talend's cloud-native data integration platform, including pre-built connectors and templates for integrating with popular cloud services such as Amazon S3, Amazon Redshift, Amazon RDS, Azure SQL Database, and Azure Data Lake Storage.
* Experience deploying **Teradata** on various cloud platforms, including AWS, Azure, and Google Cloud.
* Extending **Hive** functionality by using custom UDF’s and UDAF’s. Developed **UDFs** for encrypting and decrypting data from hive tables.
* Designed and developed interactive and visually compelling reports and dashboards using Power BI, incorporating charts, graphs, tables, and other data visualizations to present data in a clear and actionable way.
* In-depth understanding and experience with real-time data streaming technologies such as Kafka and Spark Streaming.
* Solid familiarity with many **Azure** services like virtual machines, Storage (Blob, Table, Queue Storage), networking (vNets, subnets, network security groups), Active Directory, Databases (SQL and Cosmos DB), Azure Data Studio.
* Experience in **Migrating SQL database to Azure data Lake**, Azure Synapse, Azure data lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data warehouse and migrating on-premises databases to ADLS using ADF.
* Extensive experience in administering Apache Spark-based analytics platforms, including Databricks, and ability to configure and manage Databricks clusters for optimal performance and scalability.
* Expertise in designing and implementing data pipelines on Databricks, including data ingestion, transformation, and loading from various sources to various destinations.
* Integrated **IICS** with cloud platform (AWS) to enable seamless data integration and analytics across multiple environments and Implemented real-time data integration solutions using **IICS** to enable near-instantaneous data processing and analysis for critical business applications.
* Strong understanding of data warehousing concepts and technologies, such as star and **snowflake** schemas, and ability to design and implement data warehousing solutions.
* Proficient in using **Tableau** Desktop for creating data visualizations, dashboards, and reports.
* Extensive experience in designing, developing, and implementing end-to-end data integration solutions using **Informatica PowerCenter** and various cloud services, such as AWS, Azure, and GCP.
* Extensive experience working with a variety of **NoSQL** databases, such as MongoDB, Cassandra, HBase, Amazon DynamoDB, Azure Cosmos DB and ability to design, develop, and implement database solutions that meet specific business requirements.
* **Experience with ETL tools, such as Apache Kafka and Apache NiFi, to stream data into NoSQL databases in real-time, and ability to create data pipelines and workflows that automate data processing and analytics.**
* Proficient in using **NoSQL** databases for data analytics, and ability to integrate NoSQL databases with analytical tools such as Apache Spark and Hadoop to perform complex analytics on large datasets.
* Experienced in architecting and designing solutions leveraging services like Cloud Bigquery, Cloud Data Flow, Cloud Pub OR Sub, Cloud BigTable.

# Certifications

**Microsoft Certified: Azure Data Scientist Associate**

<https://www.credly.com/badges/282024f4-c7fa-4e5a-9090-df583d69e7aa>

|  |
| --- |
| SKILLS: |

|  |  |
| --- | --- |
| Technology | Tools |
| AWS | EMR, Glue, Athena, DynamoDB, RDS, Data pipelines, Lake formation, S3, IAM, Step, Lambda, CloudFormation, EC2, ELB/CLB, MSK. |
| Azure | Data Lakes, Data Factory, Data Lake Analytics, Databricks, other azure services |
| GCP | Cloud SQL, Bigtable, Dataproc, GCE, GCS, GKE |
| Hadoop | Spark, Hive, Oozie, Sqoop, Kafka, HDFS, YARN, Zeppelin, and HBase |
| Programming languages | Java, Python, PySpark, Glue ETL, SAS |
| Database (SQL and NoSQL) | DynamoDB, HBase, Teradata, MongoDB, MYSQL, SQL SERVER 2008, PostgreSQL, Apache Cassandra, Oracle DB, Firestore, Cosmos DB, Amazon Arora |
| ETL | Informatica, Talend, Apache NiFi, MSSQL Server Integration Services (SSIS) |
| Version control | Git, SVN, Bitbucket, GitHub, SourceTree |
| Scripting languages | Shell scripting, PowerShell, Bash |
| DevOps platforms | Docker, Jenkins, Kubernetes, Ansible |
| Streaming platforms | Kafka, Confluent Kafka, Spark Streaming, Amazon Kinesis, Apache Flink, Spark Streaming, Google Cloud Dataflow, Azure Stream Analytics |
| Data Warehouses | Amazon Redshift, Azure Synapse Analytics, Google BigQuery, SAP HANA, Snowflake, Teradata, IBM Db2 |
| AI and Machine Learning | Amazon SageMaker, Google Cloud AI, Azure Machine Learning, TensorFlow, PyTorch |
| IDEs/ Editors | PyCharm, Eclipse, IntelliJ IDEA, VS code, Atom, Jupyter Notebooks, Spyder, Sublime Text, Dataiku, Zeppelin |
| Visualization Tools | Tableau, Power BI, SAS, Matplotlib, Seaborn, Bokeh, QlikView, Plotly |
| Workflow Scheduling / Management | Airflow, Apache Oozie, Apache NiFi, Google Cloud Composer, Talend, Informatica |

# Career Profile

**Client : Verizon Wireless, New Jersey (remote) May 2021 to till date**

**Position : AWS – GCP Data Engineer**

**Roles and responsibilities:**

* Experience with migrating structured, unstructured data to and from DynamoDB using AWS Data Pipeline or other ETL tools. Knowledge of best practices for data modeling, indexing, and querying in DynamoDB.
* Experience with ETL pipelines in and out of data warehouse using a combination of Python and Snowflakes Snow SQL.
* Designed and have a good understanding on the **Data modeling** (Dimensional & Relational) concepts like Star-Schema Modeling, Snowflake Schema Modeling, Fact and Dimension tables.
* Experience working with **Amazon Kinesis** for real-time data processing and integration with other AWS services.
* Designed and Implemented data pipelines using AWS services such as **Amazon Kinesis**, **Amazon Managed Streaming for Apache Kafka (MSK)** and **Spark Streaming.**
* Hands-on experience designing and implementing decoupled communication between microservices using Amazon **SQS**.
* Designed robust and scalable **Kafka clusters** on **Amazon MSK** to handle large volumes of data and ensure high availability and fault tolerance.
* Developed custom **Terraform modules** to provision data-related resources such as EMR clusters, S3 buckets, or Redshift databases.
* Designed and developed data pipelines to extract, transform, and load (ETL) data from various sources, including relational databases, flat files, and cloud-based data stores, into Power BI for reporting and analysis.
* Designed and developed PySpark applications to process and analyze large datasets from various sources, including structured and unstructured data, using distributed computing techniques.
* Implemented Airflow hooks to interact with AWS services and using Airflow macros to pass parameters between tasks. Knowledge to automate data ingestion, processing, and retrieval using **Apache Airflow**.
* Designed and implemented data integration solutions using **IICS** to connect and integrate various data sources, including on-premises systems (like Oracle), cloud platforms (like AWS) and third-party applications.
* Built and managed complex **data pipelines in IICS** to ensure efficient and reliable data extraction, transformation, and loading (ETL) processes.
* Designed and optimized data workflows in IICS for large-scale data processing, leveraging advanced features such as parallel processing, data partitioning, and error handling mechanisms.
* Experience working with various data sources, such as relational databases, flat files, XML, and JSON, and using **Informatica** PowerCenter transformations (e.g., expression, aggregator, joiner, lookup, filter) to cleanse, transform, and enrich data before loading it into target systems.
* Integrated **BigQuery** with other GCP services such as Cloud Storage, Cloud Dataflow, and Cloud Pub/Sub to create end-to-end data processing pipelines for analytics and machine learning applications. Designed and implemented data pipelines using **BigQuery**.
* Designed efficient and scalable database structures to support complex business requirements and have extensive knowledge on data modelling and schema designing. Implemented various ETL processes on data and loaded the data into various distributed databases.
* Proficient in data replication techniques, implementing strategies to ensure data consistency and availability across multiple databases and distributed systems.
* Implemented SQL querying to retrieve, manipulate and analyze large datasets along with implementing a combination of SQL, Python, Spark for handling Large datasets and streaming data for transformations.
* Designed and implemented serverless data processing pipelines using Google Cloud Functions and Cloud Storage to process and analyze large-scale datasets in real-time.
* Optimized Dataproc cluster configurations to reduce costs by 30% while maintaining high availability and performance for batch processing and real-time data ingestion.
* Developed custom Dataproc workflows using Cloud Dataproc Workflow Templates and Cloud Composer to automate data ingestion and processing tasks, resulting in a 70% reduction in manual effort and improved reliability.
* Created managed and external tables in Databricks platform using partitions for efficient data query performance.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kinesis in near real-time.
* Extensive experience designing, developing, and implementing **MongoDB** database solutions that meet specific business requirements, and proficiency in using MongoDB to manage unstructured and semi-structured data.
* Developed and executed a comprehensive data migration plan, including data mapping, schema conversion, and data validation, ensuring data integrity and minimizing downtime during the migration process.

**Environment**: Spark framework, Linux, Jira, Bitbucket, Hive, ETL, Tableau, Informatica, ETL Pipelines, Snowflake, Kafka, Spark Streaming, Apache Airflow, Apache Kafka (MSK), Agile scrum, AWS Sagemaker, MongoDB, PySpark, GCP.

**Client : CVS Health, Florham Park- NJ (remote) May 2020 to Apr 2021**

**Position : Azure Data Engineer**

**Roles and responsibilities:**

* Solid familiarity with Azure’s analytics stack - Data Lake, Data Explorer, Data Factory, Synapse, Data Bricks, HDInsight, Stream Analytics.
* Implemented Azure Data Studio to connect, manage and maintain various data sources, including Azure SQL Database, SQL Server, and other databases. Experience in creating and managing Azure Data Studio notebooks and using them to perform data analysis and visualization.
* Implemented end-to-end data engineering solutions using Azure services, including Azure Data Factory, Azure Event Hub, and Azure Stream Analytics, enabling real-time data ingestion, processing, and analysis, resulting in improved business intelligence and data-driven decision-making.
* Created train models and done experiments in Azure ML and later deployed to production.
* Implemented Azure Machine Learning Service to manage and monitor machine learning models at scale, including automating model selection, hyperparameter tuning, and deployment.
* Experience working with Azure Databricks to preprocess data and train ML models at scale using Apache Spark.
* Familiarity with Azure DevOps and GitHub to manage source code and build, test, and deploy ML models in CI/CD pipeline.
* created data workflows and ETL pipelines for ML applications implementing Azure Data Factory to orchestrate.
* Implemented Azure Synapse Analytics to enable big data processing and analytics for ML applications.
* Knowledge of integrating ADF with other Azure services such as Azure Databricks and Azure Machine Learning.
* Designed and implemented CI/CD pipelines using Azure DevOps to automate the build, test, and deployment of data engineering workflows and applications. Managed build and release pipelines for Python-based data engineering workflows, utilizing Azure DevOps tools such as YAML pipelines and Azure Artifacts.
* Implemented Azure DevOps tools such as Azure Boards and Azure Test Plans to manage project requirements, plan sprints, and conduct automated testing of data engineering solutions.
* Created and implemented Azure Databricks to develop advanced data transformation and processing workflows, leveraging its distributed computing capabilities to handle large-scale data processing and machine learning tasks, resulting in a 50% improvement in data processing performance.
* Designed in optimizing **Snowflake** performance through efficient data loading, data indexing, and data partitioning strategies.
* Utilized **Azure DevOps** to manage project requirements, plan sprints, and track progress, while utilizing Jenkins to automate the build, test, and deployment of data engineering workflows and applications.
* Designed and implemented data pipelines and ETL processes to extract, transform, and load data into **PowerBI** reports and dashboards, resulting in improved data accuracy and faster report generation times.
* Designed and developed ETL processes using Informatica PowerCenter and Informatica Data Quality (IDQ) to extract, transform, and load data from various sources into a data warehouse.
* Proficient in designing and implementing data models and machine learning pipelines using Databricks, including feature engineering, model training, and model deployment.
* Experience with integrating Databricks with other Azure services such as Azure Synapse Analytics, Azure Machine Learning, and Azure Data Lake Storage to create comprehensive data processing solutions.
* Designed and developed data integration solutions using **Azure Data Factory** to extract data from various sources, including on-premises databases, cloud-based services, and file systems.
* Implemented durable functions to handle long-running workflows and integrated with other Azure services such as Azure Blob Storage and Cosmos DB.
* Designed and optimized data pipelines using Azure Data Factory, achieving a 30% reduction in data processing time and improving overall data quality and reliability.
* Implemented data ingestion workflows using Azure Event Hub and Azure Stream Analytics, enabling near real-time data streaming and facilitating real-time analytics and monitoring of critical business metrics.
* Involved in migrating the existing v1 (Classic) Azure infrastructure into v2 (ARM), scripting and templating the whole end-to-end process as much as possible so that it is customizable for each area being migrated.
* Developed Azure Data Lakes (ADLs) and Data Lake Analytics and have good understanding on integrating ADLs with other Azure Services.

**Environment**: **Apache Spark**, **Microsoft Azure, ETL Pipelines,** SSIS, Apache Airflow, Oracle DB 10g, Cosmos DB, Teamcity, Kafka, Tableau, Data Bricks, Kubernetes, Docker, Synapse SQL, Talend Data Integration, Jenkins, Maven, Agile scrum.

**Client : General Insurance- Nashville, TN Aug 2018– Feb 2020**

**Position : AWS Data Engineer**

**Roles and responsibilities:**

* Worked with different insurance data such as claims, underwriting, finance, operations, and distributions.
* Hands-on experience on AWS components such as EMR, EC2, S3, RDS, IAM, VPC, Auto Scaling, CloudWatch, SNS, SQS, SES, Athena, Glue, Kinesis, Lambda, Redshift, DynamoDB, EBS, EFS to ensure a secure zone for an organization in AWS public cloud.
* Developed and maintained **Apache Airflow** workflows, including defining tasks, dependencies, and orchestrating data processing pipelines. Experience in deploying **Apache Airflow** on various environments, including on-premises servers and AWS.
* Designed and implemented complex data pipelines using **Airflow DAGs** to extract, transform, and load (ETL) data from various sources, resulting in streamlined and efficient data processing.
* Experience in setting up, configuring, and managing Apache Kafka clusters using **Amazon MSK**, including cluster creation, scaling, and maintenance.
* Designed and Involved in integrating AWS Step Functions with other AWS services, such as Lambda, S3, and EC2.
* Developed a framework for converting existing **Informatica PowerCenter** mappings and to PySpark Jobs. Used **Informatica** Power Center for ETL to load data from heterogeneous source systems into target databases.
* Analyzed the SQL scripts and designed them by using PySpark SQL for faster performance.
* Implement scalable and sustainable data engineering solutions using tools such as Databricks, Snowflake, Apache Spark, and Python. The data pipelines were created, maintained, and optimized as workloads move from development to production for specific use cases**.**

**Environment**: Hadoop, HDFS, Hive, **Java 1.7, Spark 1.6, SQL, HBase**, Informatica PowerCenter, UNIX Shell Scripting, MapReduce, Putty, WinSCP, Teradata, Linux, Redshift, Python, PySpark, AWS.

**Client : Symbiosys Technologies, Hyderabad Mar 2016 – Dec 2017**

**Position : Data Engineer**

**Roles and responsibilities:**

* **Extensive involvement in Designing Azure Resource Manager Template and in designing custom build steps using PowerShell, Apache Sqoop, Flume, java, MapReduce programs, hive queries, and pig scripts.**
* Generating the required reports using Oozie workflow and Hive queries for the operations team from the ingested data.
* Reporting and BI tools like Microsoft SQL Server Reporting Services (SSRS) and SAP Crystal Reports.
* Experience in ETL/Pipeline Development using tools such as Azure Databricks, Apache Spark, and Python.
* Using Pyspark to convert unstructured and semi-structured data into structured data and loading into Hive tables.
* **Importing and exporting data into HDFS, HBase, and Hive using Sqoop.**
* Experience in working on HBase with Apache phoenix as a data layer to serve the web requests to meet the SLA requirements.
* Developed Azure Databricks notebooks to apply the business transformations and used PySpark and SQL to perform data cleansing operations. Integrated the notebooks to various Databases and other services to allow easy access and analyse data from wide range of sources.
* Experience working with Delta Lake, an open-source data storage format for big data processing that provides features such as ACID transactions, schema enforcement, Data integrity, Data versioning and time travel.
* Extracted data from multiple sources including FTP, RDBMS and REST API, applied complex transformations and loaded it to ADLS.
* Used Azure DevOps & Jenkins pipelines to build and deploy different resources in Azure.
* Developed several Azure Data Factory re-usable pipelines, linked services.
* **Hands-on experience in Azure Development worked on Azure web application, App services, Azure storage, Azure SQL DB, Virtual machines, Fabric controller, Azure AD, Azure search, and notification hub.**

**Environment and Tech Stack**: SSRS, Apache Hadoop, HDFS, Pig, Hive, HBASE, **Flume, Kafka, MapReduce, Sqoop**, Spark, PySpark, Oozie, LINUX, SFDC, ODATA, and AWS, Agile scrum, PL/SQL, Azure, Databricks, NoSQL

**Client : Traingulam Technologies, HYD Oct 2013 – Jan 2016**

**Position : Data Analyst**

**Roles and responsibilities:**

* Proficient in data analysis tools such as Excel Sheets. Capable of using formulas, functions, and pivot tables to manipulate and analyse large datasets.
* Familiar with **SQL** and able to write basic queries to extract and manipulate data from relational databases.
* Skilled in using **data visualization** tools such as **Tableau** to create compelling visualizations and dashboards to communicate insights to stakeholders.
* Experience with data cleaning and preparation using **Python** libraries like **Pandas**.
* Good understanding in **statistical analysis** concepts such as hypothesis testing, regression analysis, and probability distributions.
* Implemented **Redshift**-basedData warehousing Solution.
* Familiarity with machine learning concepts and algorithms, and capable of using libraries such as Scikit-learn or TensorFlow to build basic models.
* Experienced in using **AWS Data Pipeline**, a service that makes it easy to move data between different AWS services and on-premises data sources. Created data pipelines in Data Pipeline to automate the movement and transformation of data across multiple services.
* Created SQL queries in **Athena** to explore and analyse data stored in S3, providing valuable insights to stakeholders.
* Created and managed metadata in **Glue Data Catalog** to facilitate discovery and management of data assets across the organization.

**Environments and Tech Stack**: Apache Hadoop, Java, Bash, ETL, Map Reduce, Hive, Deployment tools, Oracle 11g/10g, MySQL, Power BI, SAS, Python, PL/SQL.