**GAUTHAM DUBBA Big Data Engineer**

**Contact: +1(469)846-8345**

**G-Mail:** [saisri@mastroservices.com](mailto:saisri@mastroservices.com) **Linkedin:** [linkedin.com/in/gautham-d-b06866304](https://www.linkedin.com/in/gautham-d-b06866304)

**Professional Summary:**

* Big Data professional with Around 6+ years of combined experience in the fields of Data Applications, Big Data implementations using Apache Hadoop/Spark echo systems, Data Visualization and Analytics.
* 3.5 years’ of experience as Azure Cloud Data Engineer in Microsoft Azure Cloud technologies including Azure Data Factory(ADF), Azure Data Lake Storage(ADLS), Azure Synapse Analytics (SQL Data warehouse), Azure SQL Database, Polybase, Azure Cosmos NoSQL DB, Azure Key vaults, Azure DevOps, Big Data Technologies like Hadoop, Apache Spark and Azure Data bricks.
* 3 years of experience in Data warehouse / ETL Developer role.
* Solid experience on building ETL ingestion flows using Azure Data Factory.
* Experience in building Azure Stream Analytics ingestion spec for data ingestion which helps users to get sub second results in Real-time.
* Experience in building ETL (Azure Data Bricks) data pipelines leveraging PySpark, Spark SQL.
* Experience in building orchestration on Azure Data Factory for scheduling purposes.
* Experience working with Azure Logic APP Integration tool and experience working with Data-Warehouse like Teradata.
* Experience on Implementation of Azure log analytics providing Platform as a service.
* Experience in building the data pipeline by leveraging the Azure Data Factory and good knowledge on polybase external tables in SQL DW.
* Selecting appropriate low cost driven Azure services to design and deploy an application based on given requirements.
* Expertise on working with databases like Azure SQL DB, Azure SQL DW
* Solid programming experience on working with Python, Scala and experience working in a cross-functional AGILE Scrum team.
* Profound understanding of Partitions and Bucketing concepts in Hive and designed both Managed and External tables in Hive to optimize performance.
* Created Snowflake Schemas by normalizing the dimension tables as appropriate and creating a Sub Dimension named Demographic as a subset to the Customer Dimension.
* Hands on experience in test driven development (TDD), Behavior driven development (BDD) and acceptance test driven development (ATDD) approaches.

**Technical Skillset:**

|  |  |
| --- | --- |
| **Big Data Stack** | Hadoop Ecosystem, Map Reduce, Spark, Airflow, Nifi, Kafka, Oozie, Hadoop, Snowflake, Databricks, HBase, Hive, Pig, Sqoop. |
| **Cloud Platform** | Azure Data Factory(ADF), Azure Data Lake Storage(ADLS), Azure Synapse Analytics (SQL Data warehouse), Azure SQL Database, Polybase, Azure Cosmos NoSQL DB, Azure Key vaults, Azure DevOps, Big Data Technologies like Hadoop, Apache Spark and Azure Data bricks. |
| **Databases** | Oracle 12c/11g, MS SQL Server, My-SQL, Teradata R15/R14. |
| **IDE** | PyCharm and Eclipse |
| **Programming Languages** | Spark, Python, SCALA, SQL, HiveQL, Shell Scripting |
| **OLAP Tools** | Tableau, SSAS, Business Objects |
| **Operating System** | Windows, Unix, Sun Solaris |
| **ETL/Data warehouse Tools** | Informatica 9.6/9.1, and Tableau |
| **Data Modeling Tools** | Erwin Data Modeler, ER Studio v17 |
| **Version controls and Tools** | GIT, Maven, Bitbucket |
| **Development Methods** | Agile/Scrum, Waterfall |

**Professional Experience:**

**Sr Big Data Engineer**

**Mass Mutual, Spingfield, MA Nov 2023 to Till Date**

**Responsibilities:**

* Worked on requirements gathering, analysis and designing of the systems.
* Developed Spark programs using Scala to compare the performance of Spark with Hive and SparkSQL.
* Developed spark streaming application to consume JSON messages from Kafka and perform transformations.
* Used Spark API over Hortonworks Hadoop YARN to perform analytics on data in Hive.
* Implemented Spark using Scala and SparkSql for faster testing and processing of data.
* Involved in developing a MapReduce framework that filters bad and unnecessary records.
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs with Scala.
* Used Spark API over Hadoop YARN as execution engine for data analytics using Hive.
* Worked on building the data pipeline using Azure Service like Data Factory to load the data from Legacy, SQL server to Azure Data warehouse using Data Factories and Databricks Notebooks.
* Created dynamic pipeline to handle multiple source extracting to multiple targets; extensively used azure key vaults to configure the connections in linked services.
* Worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & AzureSQL DB).
* Create and maintain optimal data pipeline architecture in cloud Microsoft Azure using Data Factory and Azure Databricks
* Writing PySpark and spark SQL transformation in Azure Databricks to perform complex transformations for business rule implementation
* Exposed transformed data in Azure Spark Databricks platform to parquet formats for efficient data storage.
* Created and developed grafana dashboards for better analysis and monitoring.
* Analyzed function/application logs on Azure App insghts to troubleshoot data and application issues.
* Used Spark Streaming to divide streaming data into batched as an input to spark engine for batch processing.
* Analyzed function/application logs on Azure App insghts to troubleshoot data and application issues.
* Developed Python based API (RESTful Web Services) using Flask.
* Involved in Analysis, Design, Development and Production phases of the application.
* Involved in AJAX driven application by invoking web services/API and parsing the JSON response.
* Created a Git repository and added the project to GitHub.
* Constructed product-usage data aggregations using PySpark, Spark SQL and maintained in Azure Data warehouse for reporting, data science dash boarding and ad-hoc analyses.

**Environment:** Hadoop, Hive, MapReduce, Sqoop, Kafka, Spark, Yarn, Pig, PySpark, Cassandra, Oozie, Nifi, Solr, Shell Scripting, Hbase, Scala, Maven, Java, JUnit, agile methodologies, Horton works, Soap, Python, Teradata, MySQL,Azure services

**Big Data Engineer**

**AT&T, Dallas, TX Jan 2023 to Oct 2023**

**Responsibilities:**

* Processed the Web server logs by developing Multi-hop flume agents by using Avro Sink and loaded into MongoDB for further analysis, also extracted files from MongoDB through Flume and processed.
* Expert knowledge on MongoDB, NoSQL data modeling, tuning, disaster recovery backup used it for distributed storage and processing using CRUD.
* Extracted and restructured the data into MongoDB using import and export command line utility tool.
* Worked on building the data pipeline using Azure Service like Data Factory to load the data from Legacy, SQL server to Azure Data warehouse using Data Factories and Databricks Notebooks.
* Created Pipeline’s to extract data from on premises source systems to azure cloud data lake storage.
* Extensively worked on copy activities and implemented the copy behavior’s such as flatten hierarchy, preserve hierarchy and Merge hierarchy.
* Implemented Error Handling concept through copy activity.
* Exposure on Azure Data Factory activities such as Lookups, Stored procedures, if condition, for each, Set Variable, Append Variable, Get Metadata, Filter and wait.
* Configured the logic apps to handle email notification to the end users and key shareholders with the help of web services activity.
* create dynamic pipeline to handle multiple source extracting to multiple targets; extensively used azure key vaults to configure the connections in linked services.
* Worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB).
* Create and maintain optimal data pipeline architecture in cloud Microsoft Azure using Data Factory and Azure Databricks
* Writing PySpark and spark SQL transformation in Azure Databricks to perform complex transformations for business rule implementation
* Implement IOT streaming with Databricks Delta tables and Delta Lake to enable ACID transaction logging
* Exposed transformed data in Azure Spark Databricks platform to parquet formats for efficient data storage.
* Implementing Azure Logic Apps, Azure Functions, Azure Storage, and Service Bus Queues for large enterprise-level ERP Integration systems.
* Understand Business requirements, analysis and translate into Application and operational requirements.
* Designed one-time load strategy for moving large databases to Azure SQL DWH.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using Azure Data Factory and HDInsight.
* Creating Spark clusters and configuring high concurrency clusters using Azure Data bricks to speed up the preparation of high-quality data.
* Used stored procedure, lookup, execute pipeline, data flow, copy data, azure function features in ADF.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Creating Data bricks notebooks using SQL, Python and automated notebooks using jobs.
* Create and maintain optimal data pipeline architecture in cloud Microsoft Azure using Data Factory and Azure Data bricks.

**Environment:**Azure ADF, ADLS, Databricks, HDinsight, Azure SQL, dataflow jobs, copy activity, lookup activity, linked services, Hadoop, Hive, Impala, Oracle, Spark, Sqoop, Oozie, Teradata, SQL, Kafka, Zookeeper, PySpark.

**Data Engineer**

**AIG, Irving, TX Feb 2022 to Dec 2022**

**Responsibilities:**

* Developed Spark scripts by using Scala, Java as per the requirement.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Developed Scala scripts, UDFs using both Data frames/SQL/Data sets and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
* Developed Spark streaming pipeline in Java to parse JSON data and to store in Hive tables
* Worked extensively with Sqoop for importing metadata from Oracle.
* Involved in creating Hive tables and loading and analyzing data using hive queries.
* Developed Hive queries to process the data and generate the data cubes for visualizing.
* Implemented schema extraction for Parquet and Avro file Formats in Hive.
* Good experience with Talend open studio for designing ETL Jobs for Processing of data.
* Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
* Collaborated with the infrastructure, network, database, application and BI teams to ensure data quality and availability.
* Migrated an existing on-premises application to ADLS.
* Used Cloud watch logs to move application logs to blob storage and create alarms based on a few exceptions raised by applications.
* Implemented Azure provides a variety of computing and networking services to meet the needs of applications
* Writing HiveQL as per the requirements and Processing data in Spark engine and store in Hive tables.
* Importing existing datasets from Oracle to Hadoop system using SQOOP.
* Created Sqoop jobs with incremental load to populate Hive External tables.
* Writing the Spark Core Programs for processing and cleansing data thereafter load that data into Hive or HBase for further processing.
* Responsible for importing data from Postgres to HDFS, HIVE using SQOOP tool.
* Experienced in migrating HiveQL into Impala to minimize query response time.
* Implemented Avro and parquet data formats for apache Hive computations to handle custom business requirements.
* Responsible for performing extensive data validation using Hive.
* Sqoop jobs, Hive scripts were created for data ingestion from relational databases to compare with historical data.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.

**Environment**: Hadoop YARN, Spark Core, Spark SQL, Scala, Python, Java, Hive, Sqoop, Impala, Oracle, Yarn, Linux, GIT, Oozie.

**Software Developer  
Othmap, Hyderbad, India Feb 2018 to Dec 2021**

**Responsibilities**:

* Migrate data from on-premises to AWS storage buckets
* Developed a python script to transfer data from on-premises to AWS S3
* Developed a python script to hit REST API’s and extract data to AWS S3
* Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue and StepFunctions
* Created yaml files for each data source and including glue table stack creation
* Worked on a python script to extract data from Netezza databases and transfer it to AWS S3
* Developed Lambda functions and assigned IAM roles to run python scripts along with various triggers (SQS, EventBridge, SNS)
* Created a Lambda Deployment function, and configured it to receive events from S3 buckets
* Writing UNIX shell scripts to automate the jobs and scheduling cron jobs for job automation using commands with Crontab.
* Developed various Mappings with the collection of all Sources, Targets, and Transformations using Informatica Designer
* Developed Mappings using Transformations like Expression, Filter, Joiner and Lookups for better data messaging and to migrate clean and consistent data
* Designed and implemented Sqoop for the incremental job to read data from DB2 and load to Hive tables and connected to Tableau for generating interactive reports using Hive server2.
* Used Sqoop to channel data from different sources of HDFS and RDBMS.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats.
* Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
* Used Apache NiFi to copy data from local file system to HDP.
* Worked on Dimensional and Relational Data Modeling using Star and Snowflake Schemas, OLTP/OLAP system, Conceptual, Logical and Physical data modeling using Erwin.
* Automated the data processing with Oozie to automate data loading into the Hadoop Distributed File System.

**Environment:** Erwin9.8, BigData3.0, Hadoop3.0, Oracle12c, PL/SQL, Scala, Spark-SQL, PySpark, Python, kafka1.1, SAS, MDM, Oozie4.3, SSIS, T-SQL, ETL, HDFS, Cosmos, Pig0.17, Sqoop1.4, MS Access.