**BABITHA**

**Sr. Azure Data Engineer**

**PROFESSIONAL SUMMARY**

* Over 8+ years of experience in Big Data-related technologies, specializing in domains like Insurance & Finance.
* Cloud Data Engineer with expertise in Azure and AWS technologies, including Redshift, Glue, Athena, S3, Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), Azure Synapse Analytics (SQL Data warehouse), Azure SQL Database, Azure Analytical
* services, Polybase, Azure Cosmos No SQL DB, Azure Key vaults, Azure Devops, Azure HD Insight Big Data

Technologies like Hadoop, Apache Spark, and Azure Databricks.

* Proficient in Big Data technologies such as Hadoop, Spark, Azure Cosmos DB, SQL Data Warehouse, Azure DMS, Lambda, Step Function, and SQL.
* Designed Azure Cloud Architecture for hosting complex application workloads, focusing on Restful API for fully distributed systems.
* Experienced in reading Continuous JSON data lake from various source systems using Kafka into DataBricks Delta, processing files using Apache Structured Streaming, PySpark, and creating files in parquet format.
* Created manual test cases to ensure deliverables meet user requirements. Knowledgeable in Apache Hadoop ecosystem components like Spark, Cassandra, HDFS, Hive, SQOOP, and Airflow.
* Designed, coded, and tested key modules of projects using Java OOPS concepts, JSP, and worked with data formats such as CSV, JSON, and Parquet.
* Strong understanding of Data Warehousing concepts, Star schema, and Snowflake schema methodologies.
* Proficient in Angular, React, Microservices, Spring, Spring Boot, Spring Batch, Restful Web Services, Hibernate, Kubernetes, and JPA Technologies.
* Expert in implementing business rules through reusable transformations, utilizing Core Java and Executor Service for multithreading, concurrency, and memory management.
* Hands-on experience implementing Python programming and PySpark within Azure Data Factory (ADF) for building data pipelines supporting Machine Learning deployments.
* GCP experience with Data Lake, GCS, Cloud Functions, BigQuery, Databricks, and Apache Airflow (Cloud Composer).
* Extensive experience developing backend Microservices using Spring Boot, Netflix OSS (Zuul, Eureka, Ribbon, Hystrix), and following domain-driven design.
* Built efficient data pipelines for moving data between GCP and Azure using Azure Data Factory, and scripted Python for test case creation in agile sprints, also Built Power BI reports on Azure Analysis backend services for better performance compared to direct queries using GCP BigQuery.
* Utilized GCP cloud shell SDK for configuring/deploying backend services Data Proc, Storage, and Bigquery. Expertise in creating, debugging, scheduling, and monitoring jobs using Docker and Kubernetes.
* Architected multiple Data pipelines, ETL, and ELT processes for Dating section transformation in GCP, coordinating tasks within the team. Developed an Enterprise ingestion Spark framework for metadata-driven data ingestion from various sources with 100% code reuse.

**TECHNICAL SKILLS**

**Hadoop Eco-System:** Spark, Hive, Sqoop, HDFS, Oozie, Pig Azure Cloud Platform: ADFv2, BLOB Storage, ADLS, Azure SQL DB, SQL server, Azure Synapse, Azure Analytic Services, Databricks, Mapping Dataflow (MDF), Azure Data Lake (Gen1/Gen2), Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning, App Services, data lake, Logic Apps, Event Grid, Service Bus, Azure DevOps, GIT Repository Management, ARM Templates.

**Programming Languages:** Python, Scala, R, C, C++, Java, Shell Scripting Databases and Query Languages: Azure SQL Warehouse, Azure SQL DB, Micro services, Azure Cosmos No SQL DB, Teradata, Vertica, RDBMS, MySQL, Oracle, PostgreSQL, Microsoft SQL Server.

**Streaming Frameworks:** Kinesis, Kafka, Flume.

**Tools:** R Studio, PyCharm, Jupyter Notebook, IntelliJ, Eclipse, NetBeans.

**Platforms:** Linux, Windows, and OS X.

**PROFESSIONAL EXPERIENCE**

**AMFAM, Madison, Wisconsin July 2022 – Till Date.**

**Sr. Azure/Data Engineer**

**Responsibilities:**

* Used Azure Data Factory (ADF) extensively for ingesting data lake from disparate source systems of health care data and as an orchestration tool for integrating data lake from upstream to downstream systems.
* Automated jobs using different triggers (Event, Scheduled, and Tumbling) in ADF, Used Cosmos DB for storing catalog data and for event sourcing in order processing data pipelines.
* Designed and developed user-defined functions, stored procedures, and triggers for Cosmos DB also Analyzed the data flow from different sources to target to provide the corresponding design Architecture in the Azure environment.
* Took initiative and ownership to provide business solutions on time. Created High-level technical design documents and Application design documents as per the requirements and delivered clear, well-communicated, and complete design documents.
* Created DA specs and Mapping Data flow and provided the details to the developer along with HLDs. Created Build definition and Release definition for Continuous Integration and Continuous Deployment.
* Utilized Java8 features for data processing, employed Python with Django for application development, and facilitated file transfer via Azure Data Share.
* Developed front-end and back-end of the application using python on Django web framework. Created an Application Interface Document for the downstream to create a new interface to transfer and receive the files through Azure Data Share.
* Created data pipelines, data flows, and complex data transformations and manipulations using ADF and PySpark with Databricks. Containerized all the Spring Boot, Java and Node.js applications using Docker and Kubernetes.
* Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Databricks. Azure auto scaling and Application programming Interface (API) management (REST APIs).
* Configured essential Databricks clusters for both batch and continuous streaming data processing, installing the required libraries. Integrated Azure Active Directory authentication into every Cosmos DB request and demonstrated this feature to stakeholders.
* Improved performance by optimizing computing time to process the streaming data and saved cost to the company by optimizing the cluster run time.
* Performed ongoing monitoring, automation, and refinement of data engineering solutions prepare complex SQL views, stored procs in Azure SQL DW and Hyper-scale.
* Designed and developed a new solution to process the NRT data lake by using Azure stream analytics, Azure Event Hub, and Service Bus Queue.
* Demonstrated expertise in backend Micro services, focusing on RESTful API, with Spring Boot, Spring MVC, NodeJS, and Apache. Utilizing several python libraries like NumPy, Logging and Cluster.
* Experienced building data pipelines with Cloud Composer for orchestrating, Cloud Dataflow for scalable machine learning algorithms, and Cloud Data prep for data exploration.
* Created a Linked service to land the data from the SFTP location to Azure Data Lake. Created numerous pipelines in Azure using Azure Data Factory(ADF) v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks, etc.
* Had knowledge on Kibana and Elastic search to identify the Kafka message failure scenarios. Closely worked with Kafka team to set up Kafka cluster setup on the QA and Production environments. Created several Databricks Spark jobs with PySpark to perform several tables-to-table operations.
* Extensively used SQL Server Import and Export Data tool. Development of micro services with Java 8, SpringBoot, and deployment in Cloud Platforms. Created database users, logins, and permissions to set up.
* Worked with complex SQL, Stored Procedures, Triggers, and packages in large databases from various servers. Helping team members resolve any technical issues, Troubleshooting, Project Risk & Issue identification, and management also addressed resource issues, Monthly one on one, Weekly meetings.

**Environment:** Azure Cloud, Azure Data Factory (ADF v2), Java, data lake, micro services, Kafka, Azure functions Apps, Kubernetes, Azure DataLake, BLOB Storage, SQL server, data pipelines, Teradata Utilities, Windows remote desktop, API’s, lambda, UNIX Shell Scripting, AZURE Power Shell, Databricks, Python, Erwin Data Modeling Tool, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning.

**Empower Retirement, Greenwood Village, CO March 2020 – June 2022**

**Data Engineer**

**Responsibilities:**

* Used custom-developed PySpark scripts to pre-process, transform data and map to tables inside the CIF (Inmon Corporate Information Factory) data warehouse.
* Developed shell scripts of Sqoop jobs for loading periodic incremental imports of structured data from various RDMS to S3 and used Kafka to ingest real-time website traffic data to HDFS.
* As part of reverse engineering discussed issues/complex code to be resolved and translated them into Informatica logic and prepared ETL design documents.
* Experienced working with a team, leading developers, interfaced with business analysts, coordinating with management, and understanding the end-user experience.
* Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse. Micro services allow large applications to be split into smaller pieces that operate independently.
* Developed the Web Interface using spring, HTML5, Kubernetes DockerCSS3 and JavaScript.
* Experienced in building Snow pipe, In depth knowledge of Data Sharing in Snowflake Database, Schema and Table structures. Designed and implemented a fully operational production grade large scale data solution on Snowflake.
* Developed mappings in Informatica to load the data from various sources into the Data Warehouse using different transformations like Source Qualifier, Expression, Lookup, Aggregate, Update Strategy, and Joiner.
* Used Java Mailing or Messaging Service (JMS) API’s for mailing detailed notifications depending upon the success and failure once the backend process is complete React JSand also for mailing administrator of any system related problems.
* Optimized the performance of the mappings by various tests on sources, targets, and transformations.
* Scheduled the sessions to extract, transform and load data into the warehouse database on Business requirements using a scheduling tool.
* Responsible for setting up the Python REST API framework using Django and FastAPI and providing interactive python API Standard (aka. Swagger) API documentation to the other cross functional teams.
* Leveraged cloud and GPU computing technologies for automated machine learning and analytics data pipelines and experienced in Spring Boot Kafka & Micro Services.
* Extracted (Flat files, mainframe files), Transformed and Loaded data into the landing area and then into the staging area followed by integration and a semantic layer of Data Warehouse (Teradata) using Informatica mappings and complex transformations (Aggregator, Joiner, Lookup, Update Strategy, Source Qualifier, Kubernetes, Filter, Router and Expression Optimized the existing ETL pipelines by tuning SQL queries and data partition techniques.
* Created independent data marts from the existing data warehouse as per the application requirement and updated them on a bi-weekly basis.
* Decreased the Azure Data Factory(ADF) billing by pivoting from using Redshift storage to Hive tables for unpaid services and implemented various techniques like Partitioning and Bucketing over Hive tables to improve the query performance.
* Worked on Prototyping and Implementation using Java, Node JS, AWS with Slacker platform. Used Presto distributed query engine over hive tables for its high performance and low cost. Automated and validated data pipelines using Apache Airflow.

**Environment:** Azure Cloud, Azure Data Factory (ADF v2), Java, data pipelines, data lake, Azure functions Apps, API’s, Micro services, Kubernetes, Azure Data Lake, BLOB Storage, SQL server, Teradata Utilities, Kafka, Windows remote desktop, UNIX Shell Scripting, AZURE Power Shell, Databricks, Python, Erwin Data Modelling Tool, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning.

**Big Data Engineer**

**Macy's, New York, NY August 2018 to February 2020**

**Responsibilities:**

* Responsible for building scalable distributed data solution using Hadoop Cluster environment with Hortonworks distribution.
* Converted raw data with sequence data format, such as Avro and Parquet to reduce data processing time and increase data transferring efficiency through the network.
* Worked on building end to end data pipelines on Hadoop Data Platforms and also Worked on Normalization and De-normalization techniques for optimum performance in relational and dimensional databases environments.
* Designed developed and tested Extract Transform Load (ETL) applications with different types of sources.
* Created files and tuned the SQL queries in Hive Utilizing HUE. Implemented MapReduce jobs in Hive by querying the available data.
* Explored with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, pair RDD's.
* Experienced with PySpark for using Spark libraries by using Python scripting for data analysis.
* Involved in converting HiveQL into Spark transformations using Spark RDD and through Scala programming.
* Created User Defined Functions (UDF), User Defined Aggregated (UDA) Functions in Pig and Hive. Worked on building custom ETL workflows using Spark/Hive to perform data cleaning and mapping.
* Implemented Kafka Custom encoders for custom input format to load data into Kafka portions.
* Supported for the cluster, topics on the Kafka manager. Cloud formation scripting, security and resource automation.

**Environment:** Python, HDFS, MapReduce, Flume, Kafka, Zookeeper, Pig, Hive, HQL, HBase, Spark, Kafka, ETL, Web Services, Linux RedHat, Unix.

**Ceequence Technologies Hyderabad, India November 2016 – May 2018**

**Data Engineer**

**Responsibilities:**

* Configured AWS Data Pipeline for seamless data loading from S3 to Redshift, leveraging T-SQL for data extraction and transformation from diverse sources.
* Developed PySpark scripts to implement data encryption using hashing algorithms, enhancing data security.
* Proficiently built data pipelines in Azure using Airflow for ETL tasks, integrating various Azure services such as Data Proc, GCS, Cloud Functions, and Big Query.
* Micro Services were being used to gradually replace a core monolithic application while adding business features.
* Designed, developed, and tested complex database components, including stored procedures, views, and triggers, while optimizing SQL queries for performance enhancement.
* Executed ETL testing procedures, extracting, transforming, and loading data from source systems into data warehouse servers. Experienced in working with Containers with usage of Docker and Kubernetes. Professional experience in using Core Technologies Java, J2EE, Servlet, JSP, JDBC, JavaScript, HTML, XML.
* Utilized Azure Data Factory (ADF), T-SQL, Spark SQL, and U-SQL for data extraction, transformation, and loading tasks into Azure Data Storage services. Involved in supporting Logging and Monitoring of Micro Services. Having good work experience on Java, JavaScript, AJAX and CSS.
* Building/ Maintain Docker/ Kubernetes container Cluster managed by Kubernetes Linux, Bash, GIT Docker on GCP. Building database Model, APIs and Views using Python, to assemble an intelligent online arrangement.
* Expertly managed data ingestion into multiple Azure backend services like Azure Data Lake, Azure Storage, Azure SQL, and Azure Data Warehouse, followed by processing in Azure Databricks. Created comprehensive data models to correlate metrics and generates valuable insights.
* Employed AWS Data Pipeline connectors, tasks, and transformations for data extraction and loading from heterogeneous sources like Access, Excel, CSV, Oracle, and flat files.
* Conducted performance tuning of SQL queries through index optimization and execution plan enhancements to reduce runtime. Independently integrate multiple APIs and new features using Kubernetes React + GraphQL stack.
* Visualized and created dashboards of the end results using power BI.

**Environment:** MS SQL Server 2016, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), Micro Services, API’s, Management Studio (SSMS),Kubernetes, Java, Advance Excel (creating formulas, pivot tables, Hlookup, VLOOKUP, Macros), Spark, Python, ETL, PowerBI, Tableau, Hive/Hadoop, Snowflake, Power BI, AWS Data Pipeline, IBM Cognos 10.1, Data Stage, Cognos Report Studio 10.1, Cognos 8 & 10 BI, Cognos Connection, Cognos office Connection, Cognos 8.2/3/4,Data stage and Quality Stage 7.5

**Hudda Infotech Private Limited Hyderabad, India July 2015 to October 2016**

**Data Engineer**

**Responsibilities:**

* Experienced in constructing and designing multiple Data pipelines, end-to-end ETL, and ELT processes for Data ingestion and transformation in AZURE. Solid comprehension of AWS components such as EC2 and S3.
* Established a Continuous Delivery pipeline with Docker and GIT Hub. Collaborated with g-cloud function using Python to load Data into Bigquery for on-arrival CSV files in GCS bucket. Processed and loaded bound and unbound Data from Google pub/subtopic to Bigquery using cloud Dataflow with Python.
* Conducted Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python. Created both simple and complex SQL scripts to check and validate Dataflow in various applications.
* Deployed data pipelines in the cloud, including AWS and AZURE, and performed data engineering functions such as data extraction, transformation, loading, and integration in support of enterprise data infrastructures like data warehouse, operational data stores, and master data management.
* Managed data services and data backend services infrastructures with good experience in ETL concepts, building ETL solutions, and Data Modeling. Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines. Hands-on experience in architecting the ETL transformation layers and writing spark jobs for processing. Wrote Python scripts for extracting data from HTML files.
* Handled the gathering and processing of raw data at scale, involving tasks like writing scripts, web scraping, calling APIs, writing SQL queries, and developing applications. Expertise in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling, and SCD (Slowly changing dimension). Developed PL/SQL Stored Procedures, Functions, Triggers, Views, and packages. Utilized Indexing, Aggregation, and Materialized views to optimize query performance.
* Developed logistic regression models (Python) to predict subscription response rates based on customer variables such as past transactions, response to prior mailings, promotions, demographics, interests, hobbies, etc.
* Created a real-time data pipeline using Spark and processed/load bound and unbound Data from Google pub/subtopic to Bigquery using cloud Dataflow with Python. Hands-on experience in AZURE, Big Query, GCS bucket, G - cloud function, cloud dataflow, Pub/Sub cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver.

**Environment:** Azure, AWS, BigQuery, Gcs Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Docker, Kubernetes, AWS, Apache Airflow, Python, Pandas, matplotlib, seaborn library, text mining, Numpy, Scikit-learn, Heat maps, Bar charts, Line charts, ETL workflows, linear regression, multivariate regression, Python, Scala, Spark.