## **Data Engineer (AWS & Azure)**

**Name:** **Sai Teja Sri Bathina**

**Phone: (609) 945-7350**

**PROFESSIONAL SUMMARY:**

* Around 12 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, **Azure** Analytical services, Polybase, **Azure** Cosmos **NoSQL** DB, **Azure** Key vaults, **Azure** **DevOps**, **Azure** HDInsight Big Data Technologies like **Hadoop**, **Apache** **Spark** and **Azure** Data bricks.
* Big Data - **Hadoop** (MapReduce & **Hive**), **Spark** (**SQL**, Streaming), **Azure** **Cosmos DB**, **SQL** Datawarehouse, **Azure** DMS, **Azure** Data Factory, **AWS** Redshift, Athena, Lambda, Step Function and **SQL**.
* Strong knowledge in **Spark** ecosystems such as **Spark** core, **Spark** **SQL**, **Spark** Streaming libraries.
* Very Good experience working in **Azure** Cloud, **Azure** **DevOps**, **Azure** Data Factory, **Azure** Data Lake Storage, **Azure** Synapse Analytics, **Azure** Analytical services, **Azure** Cosmos NO **SQL** DB, **Azure** HD Insight Big Data Technologies (**Hadoop** and **Apache** **Spark**) and Data bricks.
* Experience in designing **Azure** Cloud Architecture and Implementation plans for hosting complex application workloads on MS **Azure**.
* Experience working in reading Continuous **JSON** data from different source system using **Kafka** into Databricks Delta and processing the files using **Apache** Structured streaming, **PySpark** and creating the files in parquet format.
* Created manual Test Cases to check that each of the deliverables meet user's requirement.
* Good knowledge in **Apache** **Hadoop** ecosystem components **Spark**, **Cassandra**, **HDFS**, **Hive**, SQOOP, **Airflow**.
* Experienced in working with different data formats CSV, **JSON** and Parquet.
* Strong in Data Warehousing concepts, Star schema and Snowflake schema methodologies, understanding Business process/requirements.
* Expert in building hierarchical and Analytical **SQL** queries that helps in reporting.
* Expert in implementing Business Rules by creating re-usable transformations like mapplets and mappings.
* Expert in using debugger in Informatica designer tool to test and fix errors in the mappings. Supported ad-hoc reporting and analytics requests with an eye for creating scalable self-service or automated solutions.
* Later Migrated applications from Django to Flask and NoSQL (DynamoDB) to SQL(**Snowflake**)
* Developed and worked on **Machine Learning algorithms** for predictive modelling.
* Architected complete scalable data pipelines, data warehouse for optimized data ingestion.
* Collaborated with data scientists and architects on several projects to create data mart as per requirement.
* Conducted complex data analysis and report on results.
* Constructed data staging layers and fast real-time systems to feed BI applications and **Machine Learning algorithms**.
* Understanding of **AWS**, **Azure** webservices and at least hands on experience working in projects. Knowledge of the software development life cycle, **Agile** **methodologies**, and test-driven development.
* Built reports for monitoring data loads into **GCP** and drive reliability at the site level.
* Develop scalable and reliable data solutions to move data across systems from multiple sources in real time (**Kafka**) as well as batch modes (**Sqoop**)
* Built Enterprise ingestion **Spark** framework to ingest data from different sources (**s3**, **Salesforce**, **Excel**, **SFTP**, **FTP** and **JDBC** Databases) which is 100% metadata driven and 100% code reuse which lets Junior developers to concentrate on core business logic rather **Spark**/**Scala** coding.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Technologies:** | HDFS, Hive, Spark, MapReduce, YARN, Spark-Core, Spark-SQL. |
| **Programming Languages:** | .NET, C/C++, HTML, SQL, PL/SQL, and Scala. |
| **Scripting Languages:** | Shell Scripting, Bash, PowerShell, Python. |
| **Operating Systems:** | UNIX, Windows, LINUX. |
| **Web technologies:** | ASP.NET, MVC Framework |
| **Cloud Technologies:** | AWS EC2, ELB, S3, Azure. |
| **Azure Stack:** | Azure Data Lake, Data factory, Azure Databricks, Azure SQL database, Azure SQL Data warehouse. |
| **Databases:** | Oracle, SQL-Server, MySQL Server, MS SQL, IBM DB2, MongoDB. |
| **Build Tools:** | ANT, Maven, Gradle, Docker and Jenkins |
| **IDE / Tools:** | Eclipse, IntelliJ, Spring Tool Suite (STS) |
| **Testing/Test Management / Defect Management tools:** | Selenium Web Driver/RC/IDE/Grid, HP Quick Test Pro (QTP), Load Runner, JIRA, Quality Center, ALM, Clear Quest, SOAP UI |
| **Version Control:** | Tortoise SVN, CVS and GIT |
| **Platforms:** | Windows, Mac, Linux and Unix. |
| **Methodologies:** | Agile, Waterfall, Test Driven Development |

**PROFESSIONAL EXPERIENCE:**

**Client: Honeywell, IL Oct 2023-Till Date**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Used **Azure** Data Factory extensively for ingesting data from disparate source systems.
* Used **Azure** Data Factory as an orchestration tool for integrating data from upstream to downstream systems.
* Migrated on-premises environment in **GCP** (Google Cloud Platform)
* 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 pipelines.
* Designed and developed user defined functions, stored procedures, triggers for **Cosmos DB**
* Analyzed the data flow from different sources to target to provide the corresponding design Architecture in **Azure** environment.
* Take 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 developer along with HLDs.
* Created Build definition and Release definition for Continuous Integration and Continuous Deployment.
* Designed **Snowflake** Schema for Data Warehouse, ODS architecture by using tools like Data Model, Erwin.
* Worked with **Talend** open studio for designing **ETL** Jobs for Processing of data.
* Developed data models and data migration strategies utilizing concepts of the **snowflake** schema.
* Created Application Interface Document for the downstream to create new interface to transfer and receive the files through **Azure** Data Share.
* Creating pipelines, data flows and complex data transformations and manipulations using ADF and **PySpark** with Databricks
* Involved in porting the existing on-premises Hive code migration to **GCP** (Google Cloud Platform) BigQuery.
* Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using **Spark** Streaming to perform streaming analytics in Data bricks.
* Created, provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
* Integrated **Azure** Active Directory authentication to every **Cosmos DB** request sent and demoed feature to Stakeholders.
* Proficient in integrating **Snowflake** with **ETL** tools like **Apache Airflow, Talend,** and Informatica to streamline data workflows.
* Experience with cloud platforms (e.g., AWS, Azure, **GCP)** and their data services.
* Improved performance by optimizing computing time to process the streaming data and saved cost to company by optimizing the cluster run time.
* Perform ongoing monitoring, automation and refinement of data engineering solutions prepare complex **SQL** views, stored procs in **Azure** **SQL** DW and Hyperscale
* Designed and developed a new solution to process the NRT data by using **Azure** stream analytics, **Azure** Event Hub and Service Bus Queue.
* Created Linked service to land the data from SFTP location to **Azure** Data Lake.
* Created numerous pipelines in **Azure** using **Azure** Data Factory v2 to get the data from disparate source systems by using different **Azure** Activities like Move &Transform, Copy, filter, for each, Databricks etc.
* Created several Databricks **Spark** jobs with **PySpark** to perform several tables to table operations.
* Extensively used **SQL Server** Import and Export Data tool.
* Created database users, logins and permissions to setup.
* Working with complex **SQL**, Stored Procedures, Triggers, and packages in large databases from various servers.
* Helping team members to resolve any technical issue, Troubleshooting, Project Risk & Issue identification and management.
* Addressing resource issues, Monthly one on one, Weekly meeting.

**Environment**: Azure Cloud, Azure Data Factory (ADF v2), Azure functions Apps, Azure Data Lake, BLOB Storage, SQL Server, Teradata Utilities, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, Erwin Data Modelling Tool, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning. GCP. Snowflake

**Client: Kennametal Apr 2019 – Sep 2023**

**Industry: Sales and Manufacturing.**

**Role: Data Engineer**

**Responsibilities:**

* Used custom developed **PySpark** scripts to pre-process, transform data and map to tables inside the CIF (Non- 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 team, lead developers, interfaced with business analysts, coordinated with management and understand the end user experience.
* Used Informatica Designer to create complex mappings using different transformations to move data to a Data Warehouse.
* Created an automation process for Distribution group which receive Inventory and sales data send activation report using **Talend** and Big query.
* 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.
* Optimized the performance of the mappings by various tests on sources, targets and transformations.
* Scheduling the sessions to extract, transform and load data in to warehouse database on Business requirements using scheduling tool.
* Extracted (Flat files, mainframe files), Transformed and Loaded data into the landing area and then into staging area followed by integration and sematic layer of Data Warehouse (**Teradata**) using Informatica mappings and complex transformations (Aggregator, Joiner, Lookup, Update Strategy, Source Qualifier, Filter, Router and Expression Optimized the existing **ETL** pipelines by tuning **SQL** queries and data partition techniques
* Created independent data marts from existing data warehouse as per the application requirement and updated them on a bi-weekly basis.
* Decreased the **Azure** 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.
* Used Presto distributed query engine over **Hive** tables for its high performance and low cost.
* Automated and validated data pipelines using **Apache** **Airflow**

**Environment:** Sqoop, Informatica, Amazon EMR/Redshift, Presto, Apache Airflow, Hive

**Client: Kennametal Aug 2016 – Mar 2019**

**Role: Data Engineer**

**Responsibilities:**

* Designed and developed a real-time matching solution for customer data ingestion.
* Worked on converting the multiple **SQL Server** and **Oracle** stored procedures into **Hadoop** using **Spark** **SQL**, **Hive**, Scala, and Java.
* Created production Data-lake that can handle transactional processing operations using **Hadoop** Eco-System.
* Developed **PySpark** and **Spark** **SQL** code to process the data in **Apache** **Spark** on Amazon EMR to perform the necessary transformations.
* Involved in validating and cleansing the data using Pig statements and hands-on experience in developing Pig MACROS.
* Analyzed dataset of 14M record count and reduced it to 1.3M by filtering out rows with duplicate customer IDs and removed outliers using boxplots and univariate algorithms.
* Worked with **Hadoop** Big Data Integration with **ETL** on performing data extract, loading, and transformation process for ERP data.
* Performed extensive exploratory data analysis using Teradata to improve the quality of the dataset and created Data Visualizations using Tableau.
* Experienced in various Python libraries like **Pandas**, One dimensional **NumPy**, and Two dimensional NumPy.
* Experienced in using **PyTorch** library and implementing natural language processing.
* Developed data visualizations in **Tableau** to display day to day accuracy of the model with newly incoming Data.
* Worked with **R** for statistical modeling like Bayesian and hypothesis test with **BAS** packages, and visualized testing results in R to delivery business insight.
* Model validation by Confusion **Matrix**, **ROC**, **AUC**, and developed diagnostic tables and graphs that demonstrated how a model can be used to improve the efficiency of the selection process.
* Presented and reported business insights by **SSRS** and **Tableau** dashboard combined with different diagrams.
* Utilized Jira as project management methodology and Git for version control to build the program.
* Reported and displayed the analysis result in the web browser with **HTML** and **JavaScript**
* Involved constructively with project teams, supported the project's goal through principle and delivered insights for team and client.

**Environment:** Hadoop, Spark SQL, Hive, Scala, Java, MS Access, SQL Server, Pig, PySpark, Tableau, Excel

**Client: Automotive Manufacturer Pvt Ltd. Apr 2014 – July 2016**

**Role: Data Analyst**

**Responsibilities:**

* Worked with Data Analyst for requirements gathering business analysis and project coordination.
* Performed migration of Reports (Crystal Reports, and **Excel**) from one domain to another domain using Import/Export Wizard.
* Wrote complex **SQL**, PL/**SQL**, Procedures, Functions, and Packages to validate data and testing process.
* Used advanced Excel formulas and functions like Pivot Tables, Lookup, If with and/index, match for data cleaning.
* Redesigned some of the previous models by adding some new entities and attributes as per the business requirements.
* Reviewed Stored Procedures for reports and wrote test queries against the source system (**SQL Server**) to match the results with the actual report against the Data mart (**Oracle**).
* Involved with data profiling for multiple sources and answered complex business questions by providing data to business users.
* Performed **SQL** validation to verify the data extracts integrity and record counts in the database tables.
* Created Schema objects like Indexes, Views, and Sequences, triggers, grants, roles, Snapshots.
* Effectively used data blending feature in **Tableau** to connect different databases like **Oracle**, MS **SQL Server**.
* Transferred data with **SAS/Access** from the databases **MS Access**, **Oracle** into SAS data sets on Windows and **UNIX**.
* Provided guidance and insight on data visualization and dashboard design best practices in Tableau.
* Performed Verification, Validation and Transformations on the Input data (Text files) before loading into target database.
* Executed data extraction programs/data profiling and analyzing data for accuracy and quality.
* Wrote complex **SQL** queries for validating the data against different kinds of reports generated by Business Objects.
* Documented designs and Transformation Rules engine for use of all the designers across the project.
* Designed and implemented basic **SQL** queries for testing and report/data validation.
* Used ad hoc queries for querying and analyzing the data.
* Performed Gap Analysis to check the compatibility of the existing system infrastructure with the new business requirements.

**Environment:** SQL, PL/SQL, Oracle9i, SAS, Business Objects, Tableau, Crystal Reports, T-SQL, SAS, UNIX, MS Access 2010

**Client: Reliance Retail. India July 2012 – Mar 2014**

**Role: Data Engineer**

**Responsibilities:**

* Worked on Hadoop eco-systems including Hive, HBase, Oozie, Pig, Zookeeper, Spark Streaming MCS (MapR Control System) and so on with MapR distribution.
* Installed and configured Hadoop MapReduce, HDFS, Developed multiple MapReduce jobs in Java for data cleaning and pre-processing.
* Built code for real time data ingestion using Java, MapR-Streams (Kafka) and STORM.
* Involved in various phases of development analyzed and developed the system going through Agile Scrum methodology.
* Involved in development of Hadoop System and improving multi-node Hadoop Cluster performance.
* Worked on analyzing Hadoop stack and different big data tools including Pig, Hive, HBase database and Sqoop.
* Developed data pipeline using flume, Sqoop and pig to extract the data from weblogs and store in HDFS
* Worked with different data sources like Avro data files, XML files, JSON files, SQL server and Oracle to load data into Hive tables.
* Used Spark to create the structured data from large amount of unstructured data from various sources.
* Implemented usage of Amazon EMR for processing Big Data across Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Performed transformations, cleaning and filtering on imported data using Hive, MapReduce, Impala and loaded final data into HDFS.
* Developed Python scripts to find vulnerabilities with SQL Queries by doing SQL injection.
* Experienced in designing and developing POC’s in Spark using Scala to compare the performance of Spark with Hive and SQL/Oracle.
* Specified the cluster size, allocating Resource pool, Distribution of Hadoop by writing the specification texts in JSON File format.
* Imported weblogs & unstructured data using the Apache Flume and stores the data in Flume channel.
* Exported event weblogs to HDFS by creating a HDFS sink which directly deposits the weblogs in HDFS.
* Used RESTful web services with MVC for parsing and processing XML data.
* Managing the OpenShift cluster, which involves scaling up and down the Amazon Web Services application nodes.
* Collaborated and communicated the results of analysis to the decision makers by presenting actionable insights by using visualization charts and dashboards in Amazon Quick Sight.
* Developed data warehouse model in snowflake for over 100 datasets using WhereScape.
* Worked on various data modeling concepts like star schema, and snowflake schema in the project.

**Environment:**Hadoop, Apache Spark, HDFS, Hive, Spark SQL, Pyspark, Python, Django, Oracle SQL, Tableau, AWS, Hadoop distribution of Horton Works, Cloudera, Pig, HBase, Linux, XML, Zookeeper, Snowflake.