|  |
| --- |
| **Venkata Pentapalli** |
| Mail: [saisri@mastroservices.com](mailto:saisri@mastroservices.com) |
| Phone: +1 469 846 8345 |

**Professional Summary:**

Having Around**11 Years** of experience in Information Technology Industry which includes around 6+ years of experience in **BigDataEcosystem** like**Azure, Data Bricks, Data Lake, Azure Data Factory**,**Hadoop, snowflake,** and **SparkEcosystems** and over 4 years of experience in Java in Developing, Implementing and maintenance of various web-based applications.

* Skilled in SDLC with Requirements Analysis, Design Specification, and Testing in Waterfall and Agile methodologies.
* Proficient in designing and managing **Azure cloud** solutions, including **DataLake**, **SynapseAnalytics, SQL Database, Databricks, and Data Factory.**
* Experienced in building end-to-end data pipelines with **Azure Data Factory, Azure Synapse analytics** and **Apache Spark**, enhancing data **ingestion** and **transformation**.
* Skilled in optimizing data models and schemas for **Azure SQL Database** and **Synapse Analytics** for improved storage and query performance.
* Expert in developing and refining **ETL** processes to enhance data quality and consistency across **Azure services**.
* Proficient in Microsoft Azure/Cloud Services and Azure data solutions.
* Knowledgeable in implementing **data security** and compliance measures within **Azure** environments to meet **GDPR** and **HIPAA** standards.
* Proficient in tuning and optimizing **Azure data services**, focusing on partitioning, indexing, and proactive monitoring.
* Experienced in setting up **CI/CD pipelines** using **Azure DevOps** and **Git** for automated data solution deployments.
* Strong in programming **with SQL, Python,** and **Scala** for data manipulation and analysis in **Azure** projects.
* Proficient in designing and developing Data Pipelines using **Python** or **Pyspark**.. Knowledgeable in **NoSQL** databases like **HBase**, **Cassandra**
* Experienced in **Python** data manipulation and analysis with various libraries.
* Skilled in **ETL** workflow management and scripting with **Python**.
* Experienced in creating reports using visualization tools and data analysis with Python.
* Proficient in programming languages and tools including Scala, Java, Python, SQL, etc.
* GoodworkingknowledgeofAmazonWebServices(**AWS**)CloudPlatformwhichincludesserviceslike**EC2, S3**, **VPC,RDS, EMR,Elastic Load Balancing** (ELB), **Cloud Front, Route 53, Elastic Beanstalk (EBS),Auto Scaling, Security Groups, EC2 Container Service (ECS), Code Commit, Code Pipeline, CodeBuild, Code Deploy, Dynamo DB, Auto Scaling, Security Groups, Lambda, RedShift, Glue, CloudWatch,CloudFormation,CloudTrail,OpsWorks,Kinesis,IAM,Data Pipeline, DMS, Aurora, ETL** and other **AWS Services.**
* Solid Excellent experience in creating cloud based solutions and architecture using **Amazon Web services (Amazon EC2, Amazon S3, Amazon RDS, EMR, Glue)** and **Microsoft Azure.**
* Hands on Experience with **AWS 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.**
* Solid Knowledge of **AWS** services like **AWS EMR, Redshift, S3, EC2, Lambda, Glue** and concepts, configuring the servers for **auto-scaling** and **elastic load balancing.**
* In-depth knowledge of **Snowflake Database**, Schema and Table structures. Define virtual warehouse sizing for **Snowflake** for different type of workloads.
* Worked on Data Migration from **Teradata to AWS Snowflake Environment using Python and BI tools like Alteryx.**
* Worked on **AWS Data Pipeline** to configure data loads from **S3** into **Redshift.**
* Familiar with cloud computing principles in **Azure**, including scalability and cost optimization.
* Excellent problem-solving ability and teamwork skills, collaborating effectively to meet project requirements.
* Developed and optimized **Spark** jobs using **Scala** for faster processing and complex data querying.
* Implemented **Spark DataFrames** on **Cloudera** for advanced **Hive data analytics**.
* Managed real-time data processing with **Spark Streaming** and **Kafka**, improving data availability and system responsiveness.
* Managed **Kafka-driven** data flows and integrated Storm topologies for enhanced preprocessing.
* Well-versed in Big Data technologies and Hadoop ecosystem.
* Created data integrations between **SQL Server** and **Hadoopclusters** using **Apache**NiFi.
* Coordinated Hadoop workflows using **Oozie** and **Zookeeper** to boost automation and system uptime.
* Administered **Hadoop** and **Vertica clusters**, supporting diverse **data warehousing** needs.
* Experienced in Data Integration and Data Warehousing using ETL tools. Strong background in Data Analysis, Cleansing, Transformation, and Data Migration with ETL tools.
* Involved in optimizing algorithms in Hadoop using Spark and related tools.
* Worked with **Hortonworks** and **Cloudera Hadoop** distributions, executing scripts and managing clusters.
* Monitored **ETL** processes and ensured data integrity in **Vertica** and **Teradatawarehouses**.
* Hands on experience on **Google Cloud Platform (GCP)** in all the bigdata products **BigQuery, Cloud Data Proc, Google Cloud Storage, Composer (Air Flow as a service)**.
* Proficient in running Impala queries and integrating **BI** tools for ad-hoc **Hadoop** data analysis.
* Applied Agile and Waterfall methodologies for software development and testing.
* Used version control tools like **CVS, GIT,** and **SVN** to support code management and team collaboration.
* Used **JIRA**, **Maven**, **MS** Build, and **Jenkins** for project tracking and **CI/CD** processes.
* Utilized Kubernetes and Docker for CI/CD system runtime
* Experienced in writing SQL and PL/SQL queries for complex data retrieval.
* Expertise in Cloudera ecosystem and data processing with various tools.
* Strong experience in writing scripts using **Python API, PySpark API** and **Spark API** for analyzing thedata.
* Experience on designing and developing **Data Pipelines** for Data Ingestion or Transformation using **Python or Pyspark.**
* Experience on **Star Schema, Snowflake** using **Sqoop** to import and Export data into **HDFS** from **RDBMS** and **vice**-versa.
* Analysed data and provided insights with **Python Pandas**
* Hands-on experience in **AWS** and Azure cloud platforms and big data tools.
* Skilled in developing microservices and lambda functions in **AWS**.
* Extensive experience in **AWS** and Azure cloud solutions and architecture.
* Familiar with workflow scheduling using various tools like Airflow, AWS Data Pipelines, etc.
* Strong skills in visualization tools like **Power BI, Tableau,** and **Excel**.
* Experienced in ingesting data into HDFS from various relational databases.
* Good knowledge and experience with both NoSQL and SQL databasesenvironment.
* Proficient in implementing and orchestrating data pipelines using Oozie and Airflow.

**Technical Skills:**

* **Big Data Ecosystem**: Hadoop, Map Reduce, Hive, YARN, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, Pig, Spark, Ambari, MongoDB, Cassandra, Storm.
* **Hadoop Distributions**: Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR and Apache
* **Cloud Technologies:** Azure SQL Database, Azure Data factory, Azure Synapse Analytics, AWS,GCP
* **Languages**: Java, Python, SQL, Scala, and JavaScript
* **No SQL Databases**: Cassandra, MongoDB and HBase
* **DB Languages:** MySQL, PL/SQL, PostgreSQL and Oracle
* **Java Technologies:** Servlets, JavaBeans, JSP, JDBC, JNDI, EJB and struts
* **Methodology:** Agile, waterfall
* **Development / Build Tools:** Eclipse, Ant, Maven, IntelliJ, JUNIT and log4J.
* **Third Party Tools**: Outline Extractor, SQL Developer, Putty, WINSCP.

**Educational Qualification:**

* **Bachelors:** Sunrise University, Computer Science and Engineering **GPA:** 3.77
* **Maters :** University of Central Missouri Big Data Analytics

**Professional Experience:**

**Client:National Veterinary Associates – Dallas, TX (Remote) Aug’22 -Present**

**Role: Sr. Azure Data Engineer**

**Responsibilities:**

* Contributed to the development of PySpark Data Frames in Azure Data bricks to read data from Data Lake or Blob storage and utilize Spark SQL context for transformation.
* Created Azure Data Factory (ADF) for copying data from Azure BLOB storage to SQL Server.
* Developed ETL pipelines in and out of data warehouse using a combination of Python and Snowflake.
* Used Snow SQL to write SQL queries against Snowflake.
* Demonstrated proficiency in programming languages like Python and Scala.
* Implement ad-hoc analysis solutions using Azure Data Lake Analytics/Store, HDInsight/Databricks.
* Create Reusable ADF pipelines to call REST APIs and consume Kafka Events.
* created and executed scalable ETL pipelines with Databricks to process and import massive datasets from different sources.
* Databricks notebooks and Databricks processing are used for automated scheduling and data processing.
* used Databricks SQL and Spark to do data analysis and produce insights. Interactive reports and dashboards were created for corporate stakeholders.
* Workflows for data processing were optimised with Databricks Delta Lake for both cost and performance.
* incorporated caching, parallel processing, and data partitioning best practices.
* Developed and implemented machine learning models using Databricks MLflow in close collaboration with data scientists.
* Helped with data preparation and feature engineering for machine learning initiatives.
* Put data quality assurance and validation procedures into place for Databricks pipelines.
* Made sure industry standards and data governance regulations were followed.
* Created tables, optimised performance, and designed and oversaw Snowflake data warehouses.
* Time travel and cloning functionalities of Snowflake were implemented to preserve previous data and expedite development processes.
* Created and oversaw data integration processes that used tools like Snow pipe, ETL/ELT frameworks, and third-party connectors to import data into Snowflake from a variety of sources.
* Guaranteed low latency and smooth data transformation and intake.
* Reduced compute costs and increased performance by optimising materialised views and sophisticated SQL queries.
* To find and fix bottlenecks, I used Snowflake's query profiling and performance monitoring tools.
* Establish strong security measures in place to safeguard sensitive data, such as encryption, role-based access control, and data masking.
* Maintained access levels, responsibilities, and permissions for users in compliance with company guidelines.
* Made use of Snowflake's data sharing features to facilitate safe data exchange and cooperation between various teams and outside partners.
* Oversaw and improved data sharing procedures to provide business intelligence and real-time analytics.
* Integration of data storage solutions in spark - especially with Azure Data Lake storage and Blob snowflake storage.
* Involved in Migrating Objects from Teradata to Snowflake and created Snow pipe for continuous data load.
* Design and Develop Data application using HDFS, Hive, Spark, Scala, Sqoop, Automic Scheduler, DB2, SQL Server and Teradata.
* Used Control-M for scheduling DataStage jobs and used Logic Apps for scheduling ADF pipelines.
* Developed data ingestion pipelines using Kafka and Flink to stream data into Druid.
* Developed a common Flink module for serializing and deserializing AVRO data by applying schema.
* Loaded data into Parquet Hive tables from Avro Hive tables.
* Worked with PowerShell scripting, Bash, YAML, Json, GIT, Rest API, and Azure Resource Management (ARM) templates to build and manage CI/CD pipelines.
* Implemented large Lambda architectures using Azure Data platform capabilities like Azure Data Lake, Azure Data Factory, HDInsight, Azure SQL Server
* Extracted, transformed, and loaded data from source systems to Azure Data Storage services using a combination of Azure Data Factory, Databricks, Pyspark, Spark SQL, and U-SQL Azure Data Lake Analytics.
* Migration of on-premises data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake and Stored (ADLS) u Azure Data Factory (ADF V1/V2).
* Created pipelines in Azure Data Factory to facilitate the extraction, transformation, and loading of data from Azure SQL, Blob storage, and Azure SQL Data Warehouse.
* Worked extensively with Azure BLOB and Data Lake storage, efficiently loading data into Azure SQL Synapse analytics (DW).
* Developed data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL.
* Developed Spark APIs to import data into HDFS from Teradata and created Hive tables.
* Develop dynamic Data Factory pipelines using parameters and trigger them as desired using events like file availability on Blob Storage, based on schedule and via Logic Apps.
* Utilized Poly base, T-SQL queries to import huge amount of data from Azure Data Lake Store to Azure SQL Data warehouse and created Azure Run book to Scale up & down Azure Analysis Services and Azure SQL Data warehouse.
* Created SQL scripts and queries to validate and import data into a SaaS platform, employing Python functions to transform data from Azure storage to Azure SQL on the Azure Databricks platform.
* Used all the functionalities of visual in Power BI like drill up, drill down, conditional formatting, tool tips.
* Develop Power BI and SSRS reports, Create SSAS Database Cubes to facilitate self-service BI.
* Implemented DAX expressions for MTD and YTD based on slicer selection.
* Integrated on on-premises data (MySQL, HBase) with cloud (Blob Storage, Azure SQL DB) and applied transformations to load back to Azure Synapse using Azure Data Factory.
* Built and published Docker container images using Azure Container Registry and deployed them into Azure Kubernetes Service (AKS).
* Developed the Python scripts to build ETL pipeline and Directed Acyclic Graph (DAG) workflows in Airflow, Apache NiFi.
* Optimized workflows by building DAGs in Apache Airflow to schedule the ETL jobs and implemented additional components in Apache Airflow like Pool, Executors, and multi-node functionality.
* Processed huge datasets by leveraging Spark Context, Spark SQL, and Spark Streaming.
* Designed and implemented a real-time data streaming solution using Azure EventHub.
* Developed Spark Streaming applications to process real-time data from various sources, such as Kafka and Azure Event Hubs.
* Built streaming ETL pipelines using Spark Streaming to extract data from various sources, transform it in real- time, and load it into a data warehouse like Azure Synapse Analytics.
* Utilized tools like Azure Databricks or HDInsight to scale out the Spark Streaming cluster as per requirements.
* Data cleaning, pre-processing and modelling using Spark and Python.
* Conducted data profiling and transformation on raw data using Python.
* Improved the query performance by transitioning log storage from Cassandra to Azure SQL Datawarehouse.
* Implemented custom-built input adapters using Spark, Hive, and Sqoop to ingest data for analytics from various sources (Snowflake, MS SQL, MongoDB) into HDFS.
* Imported data from web servers and Teradata using Sqoop, Flume and Spark Streaming API.
* Developed map-reduce jobs using Scala for compiling program code into bytecode for the JVM for data processing.
* Monitored workflows for daily incremental loads from RDBMSs (MongoDB, MS SQL, MySQL).
* Proficient in Snowflake data modeling, employing ELT methodologies through Snowpipe, and proficiently developing stored procedures.
* Implemented indexing to data ingestion using Flume sink to write directly to indexers deployed on a cluster.
* Delivered data for analytics and Business intelligence needs by managing workloads using Azure Synapse.
* Managed resources and scheduling across the cluster using Azure Kubernetes Service(AKS).
* Installed and configured Hadoop MapReduce and HDFS, developing Java and Scala MapReduce jobs for data cleaning and preprocessing.
* Developed code.js scripts to extract Apache Ranger policies from REST endpoints across different clusters and stored them in Cosmos DB.
* Designed Data Marts using dimensional data modeling techniques, including star and snowflake schemas.
* Proficient at using Alteryx for Data Blending, Data Cataloging, Data Profiling and In- Database Processing
* 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 Kafka in near real time and persists into Cassandra implementing massive data lake pipelines.
* Performed data analysis and design, and creates and maintains large, complex logical and physical data models, and metadata repositories using ERWIN.

**Environment:** Azure (Data Factory, ADF, AKS, Blob Storage), HDFS, Kubernetes, Docker, Java, Selenium ER Studio, Teradata, Oracle, Alteryx, Python, Power Bi, Hadoop, Snowflake, Spark, PySpark, Scala, Hive, SQL Server SSIS, SSRS, and SSAS, Kafka, MongoDB, MS SQL, Azure Devops, DataStage, Snowflake, Erwin.

**Client: AT&T Big Data Inc. Dallas Jan 2021 – July22**

**Role: Azure Data Engineer**

**Responsibilities:**

* Implemented Agile - Scrum Methodology for frequent changes to client requirements and following parallel development and testing.
* Performed feasibility analysis for applying CCI on multiple fact tables in the EDW by analyzing size of tables, CCI segment misalignment etc.
* Responsible for gathering requirements for the new projects and creating the data flow model of the business requirement.
* Used Python scripts to update content in the database and manipulate files.
* Created data mapping, data governance, transformation and cleansing rules involving OLTP and OLAP.
* Analyzed the source data and worked with business users and developers to develop the data model.
* Generated periodic reports based on the statistical analysis of the data from various time frame and division.
* Developed logical data model based on the requirements utilizing Erwin.
* Made logical data models and physical data models that capture current state/future state data elements and data flows.
* Automated Landing Zone and processed zone layers on Azure storage accounts by orchestrating Azure data factory pipelines.
* Conceptualized, developed, and maintained the data architecture, data models and standards for various data integration & data warehouse projects.
* Identified the dimensions along with the measures and fact on the top of OLTP source.
* Created data integration process for ETL, involving the access, manipulation, analysis, interpretation and presentation of information from both internal and secondary data sources to the business.
* Developed data mapping, data governance, transformation and cleansing rules involving OLTP and OLAP.
* Performed quality testing of converted data, identifying root cause of issues and designing / documenting proposed solutions; developed the logical and physical data model and designed the data flow.
* Authored multitude of ETL packages to extract data from heterogenous OLTP sources such as flat files, excel files, SQL Server tables to populate dimensional data mart.
* Developed pipelines in Azure data factory for migrating data from SQL server to Synapse tables.
* Worked on different load strategies like SCD 1, SCD2 and SCD 0 for loading data from different sources to synapse tables.
* Handled different types of files like JSON, fixed width, text files with header and without header, excel files and different types of delimited files in Azure data factory.
* Implemented notebooks in Azure data bricks for transforming complex files which can’t be handled in Azure Data Factory.
* Understanding all types of files and their requirements which requires processing and transformation.
* Orchestration of master pipeline by utilizing parameters and metadata driven solutions using control tables from Azure SQL DB where details of files will be stored to control the pipeline.
* Loading audit details like pipeline parameters, run id, timestamp, and pipeline details into Audit tables by using pipeline system variables and activity outputs in ADF.
* Handling code in Azure data bricks platform by utilizing spark SQL and pyspark functionalities.
* Deploying ADF code and SQL scripts from one environment to another with the help of Azure devops platform
* Checked in project related documents and scripts on Team Foundation Server.
* Developed and implemented data cleansing, data security, data profiling and data monitoring processes.
* Determined data ownership, resolved data conflict, and aligned enterprise data with an emphasis on maturing data governance practices, improving data integrity, and reducing operational risk due to data quality issues.
* Worked closely with the Enterprise Data Warehouse team and Business Intelligence Architecture team to understand repository objects that support the business requirement and process.
* Created several views on top of delta lake files in serverless SQL Pool Synapse.

**Environment:** SQL Server, Azure Data factory, Synapse, SSMS, Oracle, Azure Devops and Power BI

**Client: Nike Inc Inc, Beaverton, OR Jan 2019 –Dec 2020**

**Role: AWS/GCP Data Engineer**

**Responsibilities:**

* Involved in analyzing business requirements and prepared detailed specifications that follow project guidelines required for project development.
* Written Terraform scripts to automate AWS services which include ELB, Cloud Front distribution, RDS, EC2, database security groups, Route 53, VPC, Subnets, Security Groups, and S3 Bucket and converted existing AWS infrastructure to AWS Lambda deployed via Terraform and AWS Cloud Formation.
* Used Pyspark for data frames, ETL, Data Mapping, Transformation and Loading in complex and high-volume environment
* Utilized GitHub for version control and collaboration in a team of power engineers, ensuring coordinated development and deployment of data processing infrastructure and workflows.
* Implemented Git branching strategies to manage concurrent development efforts and streamline the integration of new features and improvements.
* Utilized Unity Catalog to create a centralized repository for power-related datasets, streamlining data discovery and fostering collaboration among power engineers and analysts.
* Defined and enforced metadata standards within Unity Catalog, ensuring consistency and accuracy of power grid data documentation.
* Implemented automated processes to update Unity Catalog entries, reflecting changes in the underlying data sources and maintaining catalog accuracy.
* Employed Terraform to automate the deployment and configuration of cloud resources, enhancing the scalability and reliability of financial data processing systems.
* Collaborated with security teams to implement Terraform best practices, ensuring compliance with financial industry regulations and security standards.
* Designed and maintained Terraform scripts for infrastructure versioning, enabling efficient tracking and management of changes.
* Implemented continuous integration and delivery (CI/CD) pipelines for Terraform code, ensuring rapid and reliable infrastructure updates.
* Process and store parquet files in the Data Lake using GCS in GCP for easy access and analysis.
* Read CSV and JSON files from Google Cloud Storage in GCP to get the information required for the client and partners using lambda functions on an event driven architecture.
* Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format
* Process the data from Kafka pipelines from topics and show the real time streaming in dashboards
* Worked on AWS Elastic Beanstalk for fast deploying of various applications developed with Java, PHP, Node.js, Python on familiar servers such as Apache.
* Exported the analysed data to the relational databases using Sqoop for visualization and to generate reports for the BI team Using Tableau.
* Developed an equivalent Spark Scala code for existing SAS code to extract summary insights on the hive tables.
* Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python
* ETL process in End-To-End Pipelines using python and GCP.
* Deposit clean parquet files into the Google Cloud Storage in GCP to provide the information for the partner and client.
* Implemented business use case in Hadoop/Hive and visualized in Tableau
* Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
* Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines
* Performed data extraction, transformation, loading, and integration in data warehouse, operational data stores and master data management
* Strong understanding of AWS components such as EC2 and S3
* Performed Data Migration to GCP
* Responsible for data services and data movement infrastructures
* Proficient in AWS Lambda, Glue, and CloudWatch and Utilized AWS Glue for ETL (Extract, Transform, Load) processes to prepare and move data from various sources into databases.
* Worked on Ingesting the data using StreamSets from various sources like JDBC to Hive by Sqoop jobs.
* Designing streaming ETL Data pipelines using Kafka as Staging area using Stream Sets.
* Worked on designing and developing the Real-Time Tax Computation Engine using Oracle, Stream Sets, Kafka, Spark Structured Streaming.
* Developing data pipelines for loading data in to MongoDB, Elastic Search and Influx database through the change data capture using Stream Sets
* Designing and developing the mappings, re-usable components (Address Validation, Parsing, Standardization, etc.) based on the business requirements using Informatica BDM and Stream Sets.
* Worked on architecting the ETL transformation layers and writing spark jobs to do the processing.
* Implemented event-driven architecture with AWS Lambda and Amazon MSK to keep databases synchronized in real time.
* Implemented data quality measures, including data validation checks and cleansing processes, to maintain data accuracy and reliability.
* Collaborated with cross-functional teams to gather data requirements and ensure data solutions aligned with business objectives.
* Developed documentation and data dictionaries using Erwin, facilitating clear communication and understanding of data models and structures.
* Aggregated daily sales team updates to send report to executives and to organize jobs running on Spark clusters
* Loaded application analytics data into data warehouse in regular intervals of time
* Designed & build infrastructure for the Google Cloud environment from scratch
* Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python
* Compiled data from various sources to perform complex analysis for actionable results
* Measured Efficiency of Hadoop/Hive environment ensuring SLA is met
* Optimized the Tensor flow Model for efficiency
* Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes
* Implemented a Continuous Delivery pipeline with Docker, and Git Hub and AWS
* Built performant, scalable ETL processes to load, cleanse and validate data
* Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies
* Collaborate with team members and stakeholders in design and development of data environment
* Preparing associated documentation for specifications, requirements, and testing

**Environment**: AWS, GCP, Bigquery, Gcs Bucket, G-Cloud Function, SAS, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Bq Command Line Utilities, Dataproc, Cloud Sql, MySQL, Posgres, Sql Server, Stream sets, Python, Scala, Spark, Hive, Spark –Sql

**Client: Macy’s, New York Oct 2017– Dec 2018**

**Role: AWS Data Engineer**

**Responsibilities:**

* Perform Informatica Cloud Services, Informatica Power Center Administration ETL strategies and ETL Informatica mapping. Setting up of Secure Agent and connect different applications and its Data Connectors for processing the different kinds of data including unstructured (logs, click streams, Shares, likes, topics etc..), semi structured (XML, JSON) and structured like RDBMS.
* Worked extensively with AWS services like EC2, S3, VPC, ELB, Auto Scaling Groups, Route 53, IAM, CloudTrail, CloudWatch, CloudFormation, CloudFront, SNS, and RDS.
* Worked on Building and implementing real-time streaming ETL pipeline using Kafka Streams API.
* Deploy and monitor scalable infrastructure on Amazon web services (AWS) and configuration management instances and Managed servers on the Amazon Web Services (AWS) platform using Ansible configuration management tools and Created instances in AWS as well as migrated data to AWS from data Center.
* Developed Python scripts to parse XML, Json files and load the data in AWS Snowflake Data warehouse.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, Parquet/Text Files into AWS Redshift.
* Used AWS Redshift, S3, Spectrum and Athena services to query large amount data stored on S3 to create a Virtual Data Lake without having to go through ETL process.
* Loaded data into S3 buckets using AWS Glue and PySpark. Involved in filtering data stored in S3 buckets using Elasticsearch and loaded data into Hive external tables.
* Involved in converting Hive/SQL queries into Spark transformations using Spark Data frames and Scala.
* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance and auto-scaling in AWS Cloud formation.
* Developed a common framework using spark to ingest data from different data sources (Teradata to S3 and S3 to Snowflake etc.,)
* Performed end- to-end Architecture & implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue and Kinesis.
* Automated data storage from streaming sources to AWS data lakes (S3, Redshift, RDS) by configuring AWS Kinesis (Data Firehose) and utilizing AWS Lambda functions for resource triggering.
* Developed micro services using AWS Lambda to make API calls for third party vendors like Melissa, Strike iron.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
* Designing ETL processes using Informatica to load data from Flat Files, Oracle, and Excel files to target Snowflake Data Warehouse database.
* Extensive Knowledge and hands-on experience implementing PaaS, IaaS, SaaS style delivery models inside the Enterprise (Data centre) and in Public Clouds using like AWS and Kubernetes etc.
* Applied required transformation using AWS Glue and loaded data back to Redshift and S3.
* Migrated data from Transactional source systems to Redshift data warehouse using spark and AWS EMR.
* Experience in analyzing and writing SQL queries to extract the data in Json format through Rest API calls with API Keys, ADMIN Keys and Query Keys and load the data into Data warehouse.
* Designed and implemented ETL pipelines between from various Relational Data Bases to the Data Warehouse using Apache Airflow.
* Implemented Spark Scripts using Spark Session, Python, Spark SQL to access hive tables into spark for faster processing of data.
* Develop framework for converting existing PowerCenter mappings and to PySpark(Python and Spark) Jobs.
* Developed various Shell Scripts for scheduling various data cleansing scripts and loading process and maintained the batch processes using Unix Shell Scripts.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster processing of data.
* Worked on Postman using HTTP requests to GET the data from RESTful API and validate the API calls.
* Developed the Pysprk code for AWS Glue jobs and for EMR.
* Created custom T-SQL procedures to read data from flat files to dump to SQL Server database using SQL Server import and export data wizard.
* Developed real-time streaming applications using PySpark, Apache Flink, Kafka, and Hive on distributed Hadoop clusters, enabling efficient and timely data processing.
* 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.
* Developed ETL python scripts for ingestion pipelines which run on AWS infrastructure setup of EMR, S3, Redshift and Lambda.
* Configured EC2 instances and configure IAM users and roles and created S3 data pipe using Boto API to load data from internal data sources.
* Executed workflow improvements in data processes using the Alteryx processing engine and NO-SQL. Utilized Alteryx to parse third-party API data for meticulous validation.
* Provided Best Practice document for Docker, Jenkins and GIT.
* Expertise in implementing DevOps culture through CI/CD tools like Repos, Code Deploy, Code Pipeline, GitHub.
* Install and configured Splunk Enterprise environment on Linux, Configured Universal and Heavy forwarder.
* Backing up AWS Postgres to S3 on daily job run on EMR using Data Frames.
* Developed server-based web traffic using RESTful API's statistical analysis tool using Flask, Pandas.
* Analyse various type of raw file like Json, Csv, Xml with Python using Pandas, Numpy etc.
* Used bug-tracking tools like Jira, confluence and version controls Git, GitLab.
* Worked on ingesting data from JSON, CSV files using spark and EMR and store the output data in Parquet file format on S3.
* Configured lambdas using YAML and JSON parameterized CFT

**Environment:** Informatica Power Center 10.x/9.x, IDQ, AWS EMR, S3, RDS, AWS Redshift, AWS Glue, Snowflake, S3, MS SQL Server, Python, Kafka, Postman, Tableau, Unix Shell Scripting, EMR, GitHub, Lambda, Boto3, DynamoDB, Python, Kinesis, Glue, Athena, Spectrum.

**Clint: Source One Solutions, Hyderabad, INDIA July 2013 – Sept 2017**

**Role: Java Developer**

**Responsibilities:**

* Extensive Involvement in Requirement Analyzes and system implementation.
* Actively involved in SDLC phases like Analyzes, Design and Development.
* Responsible for developing modules and assist in deployment as per the client’s requirements.
* Application is implemented using JSP and servlets are used for implementing Business logic.
* Developed utility and helper classes and Server-side Functionalities using servlets.
* Created DAO Classes and Written Various SQL queries to perform DML Operations on the data as per the requirements.
* Created Custom Exceptions and implemented Exception handling using Try, Catch and Finally Blocks.
* Developed user interface using JSP, JavaScript and CSS Technologies.
* Implemented User Session tracking in JSP.
* Involved in Designing DB Schema for the application.
* Implemented Complex SQL Queries, Reusable Triggers, Functions, Stored procedures using PL/SQL.
* Worked in pair programming, Code reviewing and debugging.
* Involved in Tool development, Testing and Bug Fixing.
* Performed unit testing for various modules.
* Involved in UAT and production deployments and support activities.

**Environment:** Java, J2EE, Servlets, JSP, SQL, PL/SQL, HTML, JavaScript, CSS, Eclipse, Oracle, MYSQL, IBM WebSphere, JIRA, PL/SQL.