Madan Kumar Sutapalli

Data Engineer

CONTACT

☎ 2342963324

✉ madansutapalli@gmail.com

TECHNICAL SKILLS

AWS Services: S3, EC2, EMR, Redshift, RDS, Lambda, Kinesis, SNS, SQS, AMI, IAM, Cloud formation

Hadoop Components / Big Data: HDFS, Hue, MapReduce, PIG, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, PySpark Airflow, Kafka, Snowflake Spark Components

Databases: Oracle, Microsoft SQL Server, MySQL, DB2, Teradata

Programming Languages: Java, Scala, Impala, R language, Python.

Web Servers: Apache Tomcat, WebLogic.

IDE: Eclipse, Dreamweaver

NoSQL Databases: NoSQL Database (HBase, Cassandra, Mongo DB)

Methodologies: Agile (Scrum), Waterfall, UML, Design Patterns, SDLC.

Currently Exploring: Apache Flink, Drill, Tachyon.

Cloud Services: AWS, Azure, Azure Data Factory / ETL/ELT/SSIS Azure Data Lake Storage Azure Data bricks, GCP

ETL Tools: Talend Open Studio & Talend Enterprise Platform

Reporting and ETL Tools: Tableau, Power BI, AWS GLUE, SSIS, SSRS, Informatica, Data Stage, Alteryx

EDUCATION

Kent State University, Ohio, USA

- Masters / Computer Science (May 2023 to Dec 2024)

OBJECTIVE

Experienced Data Engineer with **6+ years** of expertise in designing, implementing, and optimizing data pipelines and architectures. Adept at transforming raw data into actionable insights by leveraging advanced data processing techniques and cloud-based technologies.

PROFILE SUMMARY

- Results-driven Data Engineer with **6+ years** of experience in designing and implementing robust data solutions to drive business insights and enhance data-driven decision-making.
- Experience in using **AWS CloudFormation**, **API Gateway**, and **AWS Lambda** in automation and securing the infrastructure on **AWS**.
- **ETL pipelines** in and out of data warehouse using **Python** with **Pandas** and **Spark**.
- Proficiency in building Data pipelines and Data loading using **Azure Data Bricks** and **Azure Data Warehouse** to control the accessibility to the database. Implemented Big Data solutions using **Hadoop** technology **stack**, including **PySpark**, **Hive**, **Sqoop**, **Avro** and **Thrift**.
- Practical experience using **KTables, Global KTables,** and **KStreams** in **Apache Kafka** and Confluent environments to work with **Kafka streaming**.
- Hands-on experience with **Spark**, **Databricks**, and **Delta Lake**.
- Practical experience with **Python** and **Apache Airflow** to create, schedule, and monitor workflows. 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).**
- Experience in using different **Hadoop** eco system components such as **HDFS, YARN, MapReduce, Spark, Pig, Sqoop, Hive, Impala, and HBase, Kafka,** and **Crontab tools.** Good experience in **Agile** and **SCRUM** methodologies.
- Experience in Performance Monitoring, Security, Trouble shooting, Backup, Disaster recovery, Maintenance and Support of **Linux** systems.
- Experienced with Dimensional modelling, Data migration, Data cleansing, Data profiling, and **ETL** Processes features for data warehouses.
- Practical experience with **Python** and **Apache Airflow** to create, schedule, and monitor workflows. Experience in migrating on premise to Windows **Azure** using **Azure** Site Recovery and **Azure** backups.
- Proficient with Container systems like **Docker** and Container orchestration like **EC2** Container Service, **Kubernetes**, worked with **Terraform**.
- Expert in designing Parallel jobs using various stages like **Join, Merge, Lookup, remove duplicates, Filter, Dataset, Lookup file set, Complex flat file, Modify, Aggregator, XML.** In-depth knowledge of Data Sharing in **Snowflake** and experienced in **Snowflake Database, Schema** and **Table structures.**
- Hands-on experience in implementing, Building, and Deployment of **CI/CD** pipelines, managing projects often including tracking multiple deployments across multiple pipeline stages (**Dev, Test/QA staging,** and **production).**
- Designed and implemented a Scalable data architecture on **AWS** using **Kubernetes**, **Terraform**, and **Snowflake**, enabling seamless data integration and processing across multiple data sources.

EMPLOYMENT HISTORY

Senior Data Engineer

Worldpay - Cincinnati, Ohio, USA Jul 2024 - Present

Worldpay is an industry leading payments technology and solutions company. I design, develop, and optimize ETL/ELT pipelines to ingest, process, and transform large datasets from various sources using Azure Data Factory (ADF) and Azure Synapse Analytics. Build scalable workflows for both batch and real-time data processing.

Key Responsibilities:

- Developing **ETL pipelines** in and out of data warehouse using combination of **Python** and **Snowflake**. **SnowSQL** Writing **SQL** queries against **Snowflake**.
- Designed and **validated data models** for Snowflake and Azure Synapse, working closely with business stakeholders to align with reporting needs.
- Responsible for the design, development and maintenance of robust Web Applications as a Senior Data Engineer, improving overall data management
- Collaborated with data scientists and analysts to gather data requirements and ensure alignment with business objectives.
- Responsible for loading the data from **BDW Oracle** database, **Teradata** into **HDFS** using **Sqoop**. Implemented **AJAX**, **JSON**, and **Java script** to create interactive web screens.
- Utilized **Azure** Logic Apps to build workflows to schedule and automate batch jobs by integrating apps, **ADF pipelines**, and other services like **HTTP** requests, email triggers etc.
- Experience in using **Kafka** as a messaging system to implement real-time Streaming solutions using **Spark Streaming**
- Creating pipelines, data flows and complex data transformations and manipulations using **ADF** and **PySpark** with Databricks.
- Developed custom aggregate functions using **Spark SQL** and performed interactive querying.
- Created Data tables utilizing PyQt to display customer and policy information and add, delete, update customer records.
- Ensured data quality and accuracy with custom **SQL** and **Hive** scripts and created data visualizations using **Python** and **Tableau** for improved insights and decision-making.
- Actively Participated in all phases of the Software Development Life Cycle (**SDLC**) from implementation to deployment.
- Have used **T-SQL** for **MS SQL** Server and **ANSI SQL** extensively on disparate databases.
- Using **Azure Cluster services**, **Azure Data Factory** V2 ingested a large amount and diversity of data from diverse source systems into **Azure Data Lake Gen2**.
- Built and configured **Jenkins** slaves for parallel job execution. Installed and configured **Jenkins** for continuous integration and performed continuous deployments.
- Automated data processing workflows to enhance efficiency and reduce manual effort.
- Ensured data integrity and consistency during migration, resolving compatibility issues with **T-SQL scripting.**
- Deployed models as **Python** package, as **API** for backend integration and as services in a **Microservices** architecture with a **Kubernetes** orchestration layer for the **Dockers** containers.
- Create a new dbt Cloud workspace, which acts as the central hub for managing dbt projects.
- Developed **Kibana Dashboards** based on the Log stash data and Integrated different source and target systems into **Elasticsearch** for near real time log analysis of monitoring End to End transactions.
- Worked on **Azure Data Factory** to integrate data of both on-prem (**MY SQL**, **Cassandra**) and cloud (**Blob storage**, **Azure SQL** DB) and applied transformations to load back to **Azure Synapse**.
- Develop metrics based on **SAS scripts** on legacy system, migrating metrics to **Snowflake** (**Azure**).
- Involved in creating **Jenkins** jobs for **CI/CD** using **Git**, **Maven**, and **Bash scripting.**
- Optimized ETL pipelines in Databricks, reducing data processing times by 25% for real-time analytics. Developed Tableau dashboards that improved decision-making accuracy and reduced reporting time by 40%.
- Participated in **Agile** ceremonies, contributing to **sprint planning**, task estimation, and retrospectives.

Technologies Used: API, Azure, Azure Data Lake, Blob, Cassandra, CI/CD, Cluster, Data Factory, Docker, EC2, Elasticsearch, ETL, Factory, Git, HDFS, Hive, Java, Jenkins, JS, Kafka, Kubernetes, lake, Lake, Oracle, PySpark, Python, SAS, Snowflake, Spark, Spark SQL, Spark Streaming, SQL, Sqoop, Tableau, Teradata.

Senior Data Engineer

The Andersons - Maumee, Ohio, USA **Sep 2023 - Jun 2024**

The Andersons, Inc. is a company that operates in the U.S. ag supply chain, helping customers grow and market their crops. Designed, developed, and maintained the robust ETL/ELT data pipelines on AWS to process structured, semi-structured, and unstructured data using AWS Glue, AWS Lambda, and Apache Spark.

Key Responsibilities:

- Provisioned high availability of **AWS EC2** instances, migrated legacy systems to **AWS**, and developed **Terraform** plugins, modules, and templates for automating **AWS** infrastructure.
- Developed reusable framework to be leveraged for future migrations that automates **ETL** from **RDBMS** systems to the Data **Lake** utilizing **Spark Data Sources** and **Hive** data objects.
- Converted existing **AWS** Infrastructure to Server less architecture (**AWS Lambda**, **Kinesis**), deploying via **Terraform** and **AWS** Cloud Formation templates.
- Experience in managing and reviewing **Hadoop** log files.
- Imported real time weblogs using **Kafka** as a messaging system and ingested the data to **Spark Streaming** and did data quality checks using **Spark Streaming** and arranged bad and passable flags on the data.
- Worked with **Spark** Core, **Spark** ML, **Spark Streaming** and **Spark SQL** and **data bricks.**
- Worked with **AWS Terraform** templates in maintaining the infrastructure as code.
- Involved in the entire lifecycle of the projects including Design, Development, and Deployment, Testing and Implementation, and support.
- Optimized data infrastructure on AWS, improving scalability and reducing costs by 20%.
- **Optimized SQL queries** and data transformations, improving performance for critical business use cases.
- Used Django evolution and manual **SQL** modifications were able to modify Django models while retaining all data, while site was in production mode.
- Improved system performance by 20% through effective data retrieval techniques and database normalization
- Used **Jira** for ticketing and tracking issues and **Jenkins** for continuous integration and continuous deployment.
- Successfully managed data migration projects, including importing and exporting data to and from MongoDB, ensuring data integrity and consistency throughout the process.
- Worked on **Jenkins** pipelines to run various steps including unit, integration and static analysis tools.
- Skilled in monitoring servers using **Nagios, Cloud watch** and using **ELK Stack- Elastic search** and **Kibana.**
- Conducted query optimization and performance tuning tasks, such as query profiling, indexing, and utilizing **Snowflake's** automatic clustering to improve query response times and reduce costs.
- Developed a fully automated Continuous Integration (**CI/CD**) system using **Git**, **Jenkins**, **MySQL** and custom tools developed in **Python** and **Bash**.
- Design and Develop **ETL** Processes in **AWS Glue** to migrate Campaign data from external sources like **S3**, **ORC/Parquet/Text** Files into **AWS Redshift.**

Technologies Used: API, AWS, Cassandra, CI/CD, ETL, Git, Glue, Hive, Java, Jenkins, Jira, Kafka, lake, Lake, Lambda, MySQL, Python, RDBMS, Redshift, S3, Snowflake, Spark, Spark Core, Spark SQL, Spark Streaming, SQL

GCP Data Engineer

(HCL) Xerox - Chennai, India **Aug 2021 - May 2023**

Xerox Holdings Corporation is an American corporation that sells print and digital document products and services. Set up and managed data storage solutions using Google Cloud Storage (GCS) for raw and processed data. Implemented the data warehousing solutions using BigQuery, ensuring optimized query performance and cost efficiency.

Key Responsibilities:

- Involved in designing different components of system like **Sqoop**, **Hadoop** process involves map reduce & hive, **Spark**, **FTP** integration to down systems.

- Managed, Configured and scheduled resources across the cluster using **Azure Kubernetes Service**.
- Creating Data Studio report to review billing and usage of services to optimize the queries and contribute in cost saving measures.
- Worked on **NoSQL** Databases such as **HBase** and integrated with **PySpark** for processing and persisting real-time streaming. Experience in **GCP Dataproc, GCS, Cloud functions**, **BigQuery**.
- Optimized complex SQL queries and data transformations, reducing processing times by 30% for business-critical reports.
- Created Amazon **VPC** to create public-facing subnet for web servers with internet access, and backend databases & application servers in a private-facing subnet with no Internet access.
- Developed Databricks **ETL** pipelines using notebooks, **Spark Data frames, SPARK SQL** and **Python scripting**.
- Created detailed documentation for **ETL pipelines**, workflows, and data standards to streamline onboarding and maintenance processes.
- Good knowledge in using Cloud Shell for various tasks and deploying services.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the **Spark Databricks cluster** and Ability to apply the **spark Data Frame API** to complete Data manipulation within **spark session.**
- Used **Python** to write Data into **JSON** files for testing **Django** Websites, Created scripts for data modelling and data import and export.
- Experienced in Google Cloud components, Google container builders and **GCP** client libraries and **Cloud SDK'S.**
- Wrote data ingestion systems to pull data from traditional **RDBMS** platforms such as **Oracle** and **Teradata** and store it in **NoSQL** databases such as **MongoDB**.
- Involved in monitoring and scheduling the pipelines using Triggers in **Azure Data Factory.**
- Monitoring **BigQuery**, Dataproc and Cloud Dataflow jobs via Stack driver for all the different environments.
- Working with **GCP** cloud using in **GCP Cloud storage, Data-proc, Data Flow, Big-Query, EMR**, **S3, Glacier** and **EC2** with **EMR Cluster**
- Build a program with **Python** and **Apache Beam** and execute it in Cloud Dataflow to run Data validation between raw source file and **BigQuery** tables.
- Build data pipelines in **Airflow/Composer** for orchestrating **ETL** related jobs using different airflow operators.

Technologies Used: Airflow, Apache, Apache Beam, API, Azure, BigQuery, Cluster, EC2, EMR, ETL, Factory, GCP, HBase, JS, Kubernetes, Oracle, PySpark, Python, RDBMS, S3, SDK, Spark, SQL, Sqoop, Teradata, VPC.

Data Engineer

AstraZeneca - Chennai, India **Mar 2018 - Jul 2021**

AstraZeneca plc is a British-Swedish multinational pharmaceutical and biotechnology company. I integrated the diverse data sources including clinical trial data, patient records, genomics data, and external data providers. Maintained detailed documentation for data pipelines, workflows, and architecture.

Key Responsibilities:

- Worked on developing **ETL** processes (Data Stage Open Studio) to load data from multiple data sources to **HDFS** using **FLUME** and **SQOOP**, and performed structural modifications using **Map Reduce, HIVE.**
- The **AWS Lambda** functions were written in **Spark** with cross - functional dependencies that generated custom libraries for delivering the **Lambda** function in the cloud.
- Written custom **UDFs** in **HIVE** according to business requirements.
- Wrote **Spark** Streaming applications to consume the data from **Kafka** topics and write the processed streams to **HBase**.
- Created Databricks Job workflows which extracts data from **SQL** Server and upload the files to **SFTP** using **PySpark** and **Python**.
- Extensively worked with **Apache Avro** and **Parquet** files and converted the data from either format Parsed Semi Structured **JSON** data and converted to **Apache** Parquet using **Data Frames** in **Spark**.
- Utilized **Elasticsearch** and Kibana for indexing and visualizing the real-time analytics results, enabling stakeholders to gain actionable insights quickly.
- Involved in various phases of Software Development Lifecycle (**SDLC**) of the application, like gathering requirements, design, development, deployment, and analysis of the application.

- Developed database triggers and stored procedures using **T-SQL** cursors and tables.
- Expertise in Creating, Debugging, Scheduling and Monitoring jobs using **Airflow** for **ETL** batch processing to load into **Snowflake** for analytical processes.
- Implemented **Kubernetes** namespaces and **RBAC** (Role-Based Access Control) policies to enforce security and access controls in the data infrastructure.
- Developed **ETL** pipelines between data warehouses using a combination of **Python** and **Snowflake**, **SnowSQL**, writing **SQL** queries against **Snowflake**.
- Implementing a Continuous Delivery framework using **Jenkins**, **Maven** in multiple environments.
- Demonstrated flexibility and efficiency in hybrid work environments, balancing collaboration with onsite teams and remote execution.
- 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**.

Technologies Used: Elasticsearch, ETL, HBase, HDFS, Java, Jenkins, JS, Kafka, Kubernetes, Lambda, Oracle, PySpark, Python, S3, Snowflake, Spark, Spark Streaming, SQL, Sqoop, Teradata