Mano Vallabhan

[Mano.vallabhan@gmail.com](mailto:Mano.vallabhan@gmail.com)

469-333-0359

**Professional Summary**

* 7+ years of experience in software industry, including 5 years of experience in **Azure cloud services**, and **Data warehouse**.
* Experience in **Azure Cloud**, **Azure Data Factory**, **Azure Data Lake storage**, **Azure Synapse Analytics**, **Azure Analytical** **services**, **Azure Cosmos NO SQL DB**, **Azure Big Data Technologies** (**Hadoop** and **Apache Spark**) and **Data bricks**.
* Experience in developing, support and maintenance for the **ETL** (Extract, Transform and Load) processes using **Talend Integration Suite**.
* Experience in developing very complex mappings, reusable transformations, sessions, and workflows using **Informatica ETL** tool to extract data from various sources and load into targets.
* Proficiency in multiple databases like **MongoDB**, **Cassandra**, **MySQL**, **ORACLE**, and **MS SQL Server**.
* Experience in Developing Spark applications using **Spark - SQL** in **Databricks** for data extraction, transformation, and aggregation from multiple file formats for analyzing and transforming the data to uncover insights into the customer usage patterns.
* Used various file formats like **Avro**, **Parquet**, **Sequence**, **JSON**, **ORC**, and text for loading data, parsing, gathering, and performing transformations.
* Good experience in **Hortonworks** and **Cloudera** for Apache Hadoop distributions.
* Designed and created **Hive** external tables using shared meta-store with Static & Dynamic partitioning, bucketing, and indexing.
* Exploring with Spark improving the performance and optimization of the existing algorithms in **Hadoop** using **Spark context**, **Spark-SQL**, **Data Frame**, pair **RDD**'s.
* Extensive hands-on experience tuning spark Jobs.
* Experienced in working with structured data using **HiveQL** and optimizing Hive queries.
* Developed Spark code using Scala and Spark-SQL for faster processing of data.
* Familiarity with libraries like **PySpark**, **Numbly**, **Pandas**, **Star base**, **Matplotlib** in **Python**.
* Writing complex **SQL** queries using joins, group by, nested queries.
* Experience in **HBase** to load data using connectors and write queries using **NOSQL**.
* Experience with solid capabilities in exploratory data analysis, statistical analysis, and visualization using **R**, **Python**, **SQL**, and **Tableau**.
* Running and scheduling workflows using **Oozie** and **Zookeeper**, identifying failures and integrating, coordinating, and scheduling jobs.
* In - depth understanding of **Snowflake cloud technology**.
* Hands on experience on **Kafka** and **Flume** to load the log data from multiple sources directly into **HDFS**.
* Widely used different features of **Teradata** such as **BTEQ**, **Fast load**, **Multifood**, **SQL Assistant**, **DDL** and **DML** commands and very good understanding of **Teradata UPI** and **NUPI**, secondary indexes and join indexes.
* Having working experience with Building **RESTful web services**, and **RESTful API**.

**Education:**

* Bachelors in CS from Anna University – 2016
* Masters in BA from Texas A&M University - 2022

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Technologies** | Hadoop, Map Reduce, HDFS, Sqoop, Hive, HBase, Flume, Kafka, Yarn, Apache Spark. |
| **Databases** | MS SQL, Oracle, Blob Storage, Elasticsearch, HBase, Amazon S3, Hadoop, Google Storage |
| **Programming Languages** | Python, Pyspark, Shell script, Perl script, SQL, Java. |
| **Tools** | PyCharm, Eclipse, Visual Studio, SQL\*Plus, SQL Developer, SQL Navigator, SQL Server Management Studio, Eclipse, Postman. |
| **Version Control** | SVN, Git, GitHub, Maven |
| **Operating Systems** | Windows 10/7/XP/2000/NT/98/95, UNIX, LINUX, OS |
| **Visualization/ Reporting** | Tableau, ggplot2, matplotlib |

**USAA – San Antonio, TX Oct 2022 – Till Date**

**Sr. Azure Data Engineer**

**Responsibilities**

* Architect and implement ETL and data movement solutions using **Azure Data Factory**, **SSIS**
* Understand Business requirements, analysis and translate into Application and operational requirements.
* Designed one-time load strategy for moving large databases to **Azure SQL DWH**.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using **Azure Data Factory** and **HDInsight**.
* Created a framework to do data profiling, cleansing, automatic restart ability of batch pipeline and handling rollback strategy.
* Design and implement database solutions in **Azure SQL Data Warehouse**, **Azure SQL**
* Lead a team of six developers to migrate the application.
* Implemented masking and encryption techniques to protect sensitive data.
* Implemented **SSIS IR** to run **SSIS** packages from **ADF**.
* Developed mapping document to map columns from source to target.
* Created azure data factory (**ADF** pipelines) using **Azure blob**.
* Performed **ETL** using **Azure Data Bricks**. Migrated on-premises **Oracle ETL** process to **Azure Synapse Analytics**.
* Worked on python scripting to automate generation of scripts. **Data curation** was done using **azure data bricks**.
* Worked on **Azure data bricks**, **PySpark**, **HDInsight**, **Azure ADW** and **Hive** used to load and transform data.
* Implemented and Developing **Hive Bucketing** and **Partitioning**.
* Implemented **Kafka**, spark structured streaming for real time data ingestion.
* Developed ETL data pipelines using Sqoop, Spark, Spark SQL, Scala, and Oozie.
* Used **Azure Data Lake** as Source and pulled data using **Azure blob**.
* Used stored procedure, lookup, execute pipeline, data flow, copy data, azure function features in **ADF**.
* Worked on creating **star schema** for drilling data. Created **PySpark** procedures, functions, packages to load data.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of **Azure Data Factory**, **T-SQL**, **Spark SQL**, and **U-SQL** **Azure Data Lake Analytics**.
* Data Ingestion to one or more Azure Services - (**Azure Data Lake**, **Azure Storage**, **Azure SQL**, **Azure DW**) and processing the data in **Azure Databricks**.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the **Spark data bricks cluster**.
* Creating **Databricks notebooks** using **SQL**, **Python** and automated notebooks using jobs.
* Creating **Spark clusters** and configuring high concurrency clusters using **Azure Databricks** to speed up the preparation of high-quality data.
* Create and maintain optimal data pipeline architecture in cloud Microsoft Azure using **Data Factory** and **Azure Databricks**

**Environment**: Hadoop, Hive, Impala, Oracle, Spark, Pig, Sqoop, Oozie, Map Reduce, Teradata, SQL, Abolition, (S3, RedShift, CFT, EMR, Cloud watch), Kafka, Zookeeper, Pyspark.

**Conduent – Florham Park, NJ Sep 2021 – Sep 2022**

**Sr. Azure Data Engineer**

**Responsibilities:**

* Used **Agile Methodology** of Data Warehouse development using **Kanbanize**.
* Developed data pipeline using **Spark**, **Hive** and **HBase** to ingest customer behavioral data and financial histories into **Hadoop cluster** for analysis.
* Working experience on **Azure Databricks** cloud to organize the data into notebooks and making it easy to visualize data using dashboards.
* Performed **ETL** on data from different source systems to Azure Data Storage services using a combination of **Azure Data Factory**, **T-SQL**, **Spark SQL**, and **U-SQL Azure Data Lake Analytics**. Data Ingestion to one or more Azure Services - (**Azure Data Lake**, **Azure Storage**, **Azure SQL**, **Azure DW**) and processing the data in **Azure Databricks**.
* Worked on managing the **Spark Databricks Azure**
* Implemented data ingestion from various source systems using **Sqoop** and **Pyspark**.
* Hands on experience implementing **Spark** and **Hive** jobs performance tuning.
* KS by proper troubleshooting, estimation, and monitoring of the clusters.
* Performed Data Aggregation, Validation and on **Azure HDInsight** using spark scripts written in **Python**.
* Performed monitoring and management of the Hadoop cluster by using **Azure HDInsight**.
* Involved in extraction, transformation and loading of data directly from different source systems (flat files/Excel/Oracle/SQL) using SAS/SQL, SAS/macros.
* Generated **PL/SQL** scripts for data manipulation, validation, and materialized views for remote instances.
* Created partitioned tables in **Hive**, also designed a data warehouse using Hive external tables and created hive queries for analysis.
* Created and modified several database objects such as Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions and Triggers using **SQL** and **PL/SQL**.
* Created large datasets by combining individual datasets using various inner and outer joins in **SAS/SQL** and dataset sorting and merging techniques using **SAS/Base**.
* Extensively worked on **Shell scripts** for running SAS programs in batch mode on **UNIX**.
* Wrote Python scripts to parse **XML documents** and load the data in database.
* Used **Hive**, **Impala** and **Sqoop** utilities and **Oozie** workflows for data extraction and data loading.
* Created **HBase** tables to store various data formats of data coming from different sources.
* Responsible for importing log files from various sources into **HDFS** using **Flume**.
* Responsible for translating business and data requirements into logical data models in support Enterprise data models, **ODS**, **OLAP**, **OLTP** and **Operational data structures**.
* Created **SSIS** packages to migrate data from heterogeneous sources such as **MS Excel**, **Flat** files, and **CVS** files.
* Provided thought leadership for architecture and the design of Big Data Analytics solutions for customers, actively drive **Proof of Concept** (POC) and **Proof of Technology** (POT) evaluations and to implement a Big Data solution.

**Environment**: ADF, Databricks and ADL Spark, Hive, HBase, Sqoop, Flume, ADF, Blob, cosmos DB, MapReduce, HDFS, Cloudera, SQL, Apache Kafka, Azure, Python, power BI, Unix, SQL Server.

**Commvault Apr 2019 – July 2021**

**AWS/Azure Data Engineer**

**Responsibilities:**

* Creating and sustaining an optimal **Data pipeline architecture**.
* Imported data from **AWS S3** into **Spark RDD**, performed transformations and actions on RDD’s.
* Designed, built, and maintained data integration programs in a **Hadoop** and **RDBMS** environment, working with both traditional and non-traditional source systems, as well as RDBMS and **NoSQL** data stores for data access and analysis, Extensive usage with Python, including the creation of a custom ingest framework.
* Migrated an existing on-premises application to **AWS**. Used AWS services like **EC2** and **S3** for datasets processing and storage. Maintained the Hadoop cluster on **AWS EMR**.
* Installed application on **AWS EC2** instances, configured the storage on **S3 buckets**, and worked closely with **AWS** **EC2** infrastructure teams to troubleshoot complex issues.
* Design and develop ETL processes on **AWS Glue** to migrate campaign data from external sources like **S3**, **ORC/Parquet/Text** Files into **AWS Redshift**.
* Designed and developed functionality to get JSON document from **MongoDB** document store and send it to the client using RESTful web service. Implemented a Data interface to get information of customers using **REST API** and pre-process data using **MapReduce** and store it into **HDFS**.
* Installed and designed with Apache bigdata Hadoop components like **HDFS**, **MapReduce**, **YARN**, **Hive**,   
  **HBase**, **Sqoop**, **Pig**, **Ambari** and **NiFi**.
* Created the infrastructure needed for optimal data extraction, transformation, and loading from a wide range of data sources.
* Used Spark Streaming to stream data from external sources using **Kafka** service and migrated an existing on-premises application to **AWS**. Used AWS services like **EC2** for processing datasets and **S3** storing small datasets.
* Responsible for loading data from the internal server and the **Snowflake** data warehouse into S3 buckets.
* Developed **AWS Athena** to ingest structured data from S3 into various systems such as **RedShift** or to generate reports.
* Created Pods and controlled them using **Kubernetes**, using **Jenkins** pipelines to push all micro service builds to the Docker registry and then deploy to Kubernetes.
* Constructing Kubernetes container clusters on **GCP**, operated by **Linux**, **Bash**, **GIT**, and **Docker** (Google Cloud Platform). To develop, evaluate, and deploy, the **CI/CD** framework used **Kubernetes** &**Docker** as the runtime environment.
* Used the **Spark API** to analyze Hive data in conjunction with the **EMR Cluster Hadoop Yarn**.
* AWS Cloud Formation templates were designed to create **VPCs**, **subnets**, and **NAT** to ensure the successful deployment of Web applications and database templates.
* Proficient in **NiFi** data handling, **NiFi** **Registry** and **Version Controlling**.
* Designing and Creating **Azure Data Factory** (ADF) extensively for ingesting data from different source systems like relational and non-relational to meet business functional requirements.
* Implemented Continuous Integration using **Jenkins**, which tracks the source code changes.
* Expertise with AWS databases such as **RDS (Aurora)**, **Redshift**, **DynamoDB**, and **Elastic Cache** (**Memcached** & **Redis**).
* Implemented end-to-end data pipeline using **FTP Adaptor**, **Spark**, **Hive**, and **Impala**.
* Worked in dealing with both **agile** and **waterfall** methods in a fast-paced manner.
* Created external tables with partitions using **AWS Athena** and **Redshift**.
* Worked with **RDBMS** including **Oracle/ DB2, SQL** Server, **PostgreSQL 9.x**, **MS Access**, and **Teradata** for faster access to data on **HDFS**.
* Generated scheduled reports for **Kibana** dashboards and visualizations.
* Worked in **Agile** development environment using **Kanban** methodology.

**Tetrasoft – Hyderabad, India July 2017 – Mar 2019**

**Data Engineer**

**Responsibilities:**

* Anchor artifacts for multiple milestones (application design, code development, testing, and deployment) in software lifecycle.
* Develop **Apache Strom** program to consume the Alarms in real time streaming from **Kafka** and enrich the alarm and pass it to **EEIM** Application.
* Creating rules Engine in Apache Strom to categorize the alarm into Detection, Interrogation & Association types before processing of alarms.
* Responsible for developing **EEIM** Application as Apache Maven project and commit to code to **GIT**.
* Analyze the Alarms and enhance the **EEIM** Application using Apache Strom to predict the root cause of the alarm and exact device where the network failure has happened.
* Accumulate the **EEIM** Alarm data to the **NoSQL** database called **Mongo DB** and retrieve it from Mongo DB when necessary.
* Build Fiber to The Neighborhood or Node **(FTTN)** Topology and Fiber to The Premises **(FTTP)** Topology using **Apache Spark** and **Apache Hive**.
* Process the system logs using log stash tool and store to elastic search and create dashboard using **Kibana**.
* Regularly tune performance of **Hive** queries to improve data processing and retrieval.
* Provide technical support for debugging, code fix, platform issues, missing data points, unreliable data source connections and big data transit issues.
* Developed Java and Python applications to call the external **REST API**s to retrieve weather, traffic, geocode information.
* Working Experience on **Azure Databricks** cloud to organize the data into notebooks and making it easy to visualize data using dashboards.
* Worked on managing the **Spark Databricks** by proper troubleshooting, estimation, and monitoring of the clusters.
* Performed Data Aggregation, Validation and on **Azure HDInsight** using spark scripts written in **Python**.
* Performed monitoring and management of the **Hadoop** cluster by using **Azure HDInsight**.
* Worked with **Jira**, **Bit Bucket,** and source control systems like **Git** and **SVN** and development tools like **Jenkins**, **Artifactory**.

**Environment**: PySpark, MapReduce, HDFS, Sqoop, flume, Kafka, Hive, Pig, HBase, SQL, Shell Scripting, Eclipse, SQL Developer, Git, SVN, JIRA, Unix.

**S&P Global – Hyderabad, India Jun 2016 – Jun 2017**

**Big Data Engineer**

**Responsibilities:**

* Involved in Requirement gathering, Business Analysis and translated business requirements into technical design in **Hadoop** and **Big Data**.
* Involved in **SQOOP** implementation which helps in loading data from various **RDBMS** sources to Hadoop systems and vice versa.
* Developed Python scripts to extract the data from the web server output files to load into **HDFS**.
* Written a python script which automates to launch the **EMR cluster** and configures the Hadoop applications.
* Extensively worked with **Avro** and **Parquet** files and converted the data from either format **Parsed Semi Structured JSON** data and converted to Parquet using **Data Frames** in **PySpark**.
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions, Documented the systems processes and procedures for future references.
* Involved in Configuring **Hadoop cluster** and load balancing across the nodes.
* Involved in Hadoop installation, Commissioning, Decommissioning, Balancing, Troubleshooting, Monitoring and, debugging Configuration of multiple nodes using **Hortonworks** platform.
* Involved in working with Spark on top of **Yarn/MRv2** for interactive and Batch Analysis.
* Involved in managing and monitoring Hadoop cluster using **Cloudera Manager**.
* Used **Python** and **Shell** scripting to build pipelines.
* Developed data pipeline using **Sqoop**, **HQL**, **Spark** and **Kafka** to ingest Enterprise message delivery data into **HDFS**. F
* Developed workflow in **Oozie** also in Airflow to automate the tasks of loading data into **HDFS** and pre-processing with **Hive**.
* Integrated Hadoop into traditional **ETL**, accelerating the extraction, transformation, and loading of massive semi structured and unstructured data. Loaded unstructured data into **Hadoop distributed File System** (HDFS).
* Created **HIVE** Tables with dynamic and static partitioning including buckets for efficiency. Also created external tables in HIVE for staging purposes.
* Loaded HIVE tables with data, wrote hive queries which run on **MapReduce** and Created customized BI tool for manager teams that perform query analytics using **HiveQL**.
* Aggregated **RDD**s based on the business requirements and converted RDDs into Data frames saved as temporary hive tables for intermediate processing and stored in **HBase/Cassandra** and **RDBM**s.

**Environment:** Hadoop 3.0, Hive 2.1, J2EE, JDBC, Pig 0.16, HBase 1.1, Sqoop, NoSQL, Impala, Java, Spring, MVC, XML, Spark 1.9, PL/SQL, HDFS, JSON, Hibernate, Bootstrap, jQuery.