ANANDU SANIL NAIR
Data Engineer
Mail: anandunair284@gmail.com
Phone: 716 939 9370

About Me

Reliable Data Engineer with over 8+ years of experience in data warehousing, data engineering, feature engineering, big data, ETL/ELT, and business intelligence. As a Data architect and engineer, specialize in AWS, Azure frameworks, Cloudera, Hadoop Eco system, Python Spark/Py Spark/Scala, Data bricks, Hive, Redshift, Snowflake, relational databases, tools like Tableau, Airflow, DBT, Presto/Athena, Data DevOps Frameworks/Pipelines with strong Programming/Scripting Skills. Expert data sets and conducting performance tuning.

Key Skills

Azure | Hadoop | Python | Aws| Azure Data bricks | Power BI | SSIS | No SQL | C#

- SQL Server / Azure SQL Database
- Azure Synapse Analytics
- Performance Tuning & Optimization
- Data modeling, OLTP/OLAP
- Azure Data Factory / ETL / ELT / SSIS
- Azure Data Lake Storage
- Azure Databricks
- SSRS / Power BI / Tableau

SUMMARY

- Data Engineer with experience in building data solutions using **SQL Server**, MSBI **AWS** Azure Cloud.
- Experience in **Azure** Cloud, **Azure** Data Factory, **Azure** Data Lake Storage, **Azure** Synapse Analytics, **Azure** Analytical services, **Azure** Cosmos NO SQL DB, and **Data bricks**.
- Well versed with Agile with Scrum, Waterfall Model and Test-driven Development (TDD) methodologies.
- Extensive working experience in implementation and operations of data governance, data strategy, data management and solutions.
- Experience on Migrating **SQL** database to **Azure** Data Lake, **Azure** Data Lake Analytics, **Azure** SQL Database, Data Bricks and **Azure** SQL Data warehouse and Controlling and granting database access and Migrating On premise databases to **Azure** Data lake store using **Azure** Data factory.
- Experience with an in-depth level of understanding in the strategy and practical implementation of AWS Cloud-Specific technologies including **EC2, EBS, S3, VPC, RDS, SES, ELB, EMR, ECS**, Cloud Front, Cloud Formation, **Elastic Cache, Cloud Watch, Red Shift, Lambda**, SNS, Dynamo DB, Kinesis.
- Hands on experience on AWS cloud services (**VPC, EC2, S3, RDS, Redshift, Data Pipeline, EMR, DynamoDB, Workspaces, Lambda, Kinesis, RDS, SNS, SQS**)
- Proficient in SQLite, **MySQL** and **SQL** databases with Python.
- Practical understanding of the **Data modeling** (Dimensional & Relational) concepts like Star - Schema modeling, **Snowflake** Schema Modeling, Fact and Dimension tables.
- Experience in handling **python** and **spark** context when writing **PySpark** programs for **ETL**.
- Strong knowledge in data visualization using **Power BI** and **Tableau**.
- Hands in experience on NoSQL database like **Snowflake**, **HBase**, **Cassandra** and **MongoDB**.
- Experience in building and architecting multiple **Data pipelines**, end to end ETL and ELT process for Data ingestion and transformation in cloud and coordinate task among the team.
- Experience with **Apache** Spark ecosystem using **Spark-Core**, **SQL**, **Data Frames** and RDD's.

- Experienced in **data manipulation** using python.
- Proficient in installing, configuring and using Apache Hadoop ecosystems such as **MapReduce, Hive, Pig, Flume, Yarn, HBase, Sqoop, Spark, Storm, Kafka, Oozie, and Zookeeper.**
- Strong experience on designing **data pipelines** such as **Data Ingestion**, **Data Processing** (Transformations, enrichment and aggregations) and Reporting.
- Experienced in integrating **Kafka** with **Spark** streaming for high speed data processing.
- Experienced in implementing **Azure** data solutions, provisioning storage account, **Azure** Data Factory, **Azure** Databricks, **Azure** Blob Storage, **Azure** Synapse and **Azure** Cosmos DB.
- Skilled in debugging and optimizing **SQL** code with standard techniques.
- Demonstrated expert level technical capabilities in areas of Azure Batch and Interactive solutions, Azure Machine learning solutions and operationalizing end to end Azure Cloud Analytics solutions.
- Proficient in using ETL (SSIS) and **Azure Data Pipeline** to develop jobs for extracting, cleansing, transforming, and loading data into data warehouse.
- Merit of utilizing **Azure BLOB storage**, **Azure SQL Server, and Azure SQL data warehouse** to manage BI activities for Cloud.
- An adaptable individual with commendable communication, negotiation, and presentation skills.
- Excellent interpersonal and communication skills, creative, research-minded, technically competent and result-oriented with problem solving and leadership skills.
- Ability to work effectively in cross-functional team environments, excellent communication, and interpersonal skills.

Technical Skills:

Hadoop Components	HDFS, Hue, MapReduce, PIG, Hive, HCatalog, Hbase, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, Kerberos, pyspark.
Spark Components	Apache Spark, Data Frames, Spark SQL, Spark, YARN, Pair RDDs
Web Technologies / Other components	J2EE, XML, Log4j, HTML, XML, CSS, JavaScript,
Server Side Scripting	UNIX Shell, Power Shell Python Scripting (Boto3)
Databases	Oracle, Microsoft SQL Server, MySQL, DB2, Teradata, snowflake
Programming Languages	Java, Scala, Impala, Python.
Web Servers	Apache Tomcat, WebLogic.
IDE	Eclipse, Dreamweaver
OS/Platforms	Windows, Linux (All major distributions), Unix, CENTOS
NoSQL Databases	Hbase, MongoDB.
Methodologies	Agile (Scrum), Waterfall, UML, Design Patterns, SDLC.
Currently Exploring	Apache, Flink, Drill, Tachyon.
Cloud Services	AWS, Azure
AWS Services	S3, EC2, EMR, Redshift, RDS, Glue, Lambda, Kinesis, SNS, SQS,AMI, IAM, Cloud formation
Azure	● Azure Data Factory / ETL / ELT / SSIS ● Azure Data Lake Storage ● Azure Databricks
ETL Tools	Talend Open Studio & Talend Enterprise Platform

Machine Learning	Regression Methods: Linear, Multiple, Polynomial, Decision trees.

EDUCATION

Masters in Data Science, University of Buffalo, The State University of New York,NY - 2017
Bachelors in Computer Science, Pillai College of Engineering,Ind- 2014

PROFESSIONAL EXPERIENCE

AbbVie (Chicago, Illinois) **Mar 2021 – till date**
Role: Sr.Data Engineer

Responsibilities:

- Understood the current Production state of application and determine the impact of new implementation on existing business processes.
- Part of the Agile Team and work on weeks sprints, daily sprint Status, sprint demo preparation and stakeholder demo and signoff.
- Worked collaboratively with all levels of business stakeholders to architect, implement and test Big Data based analytical solution from disparate sources.
- Proposed architectures considering cost/spend in **Azure** and develop recommendations to right-size data infrastructure.
- Traced and catalogue data processes, transformation logic and manual adjustments to identify data governance issues.
- Participated in the Data Governance working group sessions to create Data Governance Policies.
- Analyzed, designed and built Modern data solutions using **Azure** PaaS service to support visualization of data.
- Built Complex distributed systems involving huge amount **data handling**, collecting metrics building **data pipeline, and Analytics.**
- Developed pipelines to move the data from **Azure** blob storage/file share to **Azure** Sql data warehouse and blob.
- Developed Spark applications using **Pyspark** and **Spark-SQL** for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Continuously monitored and managed **data pipeline (CI/CD)** performance alongside applications from a single console with Azure Monitor.
- Ingested data into HDFS using Sqoop and scheduled an incremental load to HDFS.
- Worked with Hadoop infrastructure to storage data in HDFS storage and use **HIVE SQL** to migrate underlying **SQL** codebase in **Azure**.
- Uploaded streaming data from **Kafka** to HDFS, **HBase** and Hive by integrating with storm.
- Developed JSON Scripts for deploying the Pipeline in **Azure** Data Factory (ADF) that process the data using the **SQL** Activity. Worked on Cluster co-ordination services through Zookeeper.
- Monitored workload, job performance and capacity planning using **Cloudera** Manager.
- Involved in build applications using Maven and integrated with CI servers like Jenkins to build jobs.
- Exported the analyzed data to the RDBMS using Sqoop for to generate reports for the BI team.
- Created the cube in Talend to create different types of aggregation in the data and also to visualize them.
- Created Build and Release for multiple projects (modules) in production environment using Visual Studio Team Services (VSTS).
- Published Power BI Reports in the required originations and Made Power BI Dashboards available in Web clients and mobile apps.

Environment: Hadoop 3.3, HDFS, Spark 3.2, PySpark, Azure, Power BI, Sqoop, Zookeeper, ADF, Hive, Sqoop, Maven, JSON, Kafka

Baker Hughes (Houston, TX) **May 2019 - Feb 2021**
Role: Data Engineer

Responsibilities:

- Designed and built **Spark**/**Pyspark** based **ETL** pipelines for migration of credit card transactions, account and customer data into enterprise Hadoop Data Lake. Developed strategies in handling large datasets using partitions, **Spark SQL**, broadcast joins and performance tuning.
- Built and implemented performant data pipelines using **Apache** Spark on **AWS EMR**. Performed maintenance of data integration programs into **Hadoop** and RDBMS environments from both structured and semi- structured data source systems.
- Developed a 16-node cluster in designing the Data Lake with the Hortonworks distribution
- Developed performance tuning on existing Hive queries and UDF's to analyze the data. Used Pig to analyze datasets and perform transformation according to requirements.
- Supervised on data profiling and data validation to ensure the accuracy of the data between the source and the target systems. Performed job scheduling and monitoring using **Auto sys** and quality testing using **ALM**
- Worked on building of **Tableau** desktop reports and dashboards to report customer data.
- Built and published customized interactive Tableau reports and dashboards along with data refresh scheduling using Tableau Desktop.
- Developed store procedures/views in **Snowflake** and used in AWS glue for loading Dimensions and Facts.
- Wrote Python scripts to manage **AWS** resources from API calls using BOTO SDK and worked with **AWS** CLI.
- **Snowflake** - data warehouse to consume the data from C3 Platform.
- Involved in S3 event notifications, an SNS topic, an SQS queue, and a Lambda function sending a message to the Slack channel.
- Involved in designing the **Data pipeline** from end-to-end, to ingest data into the **Data Lake**.
- Transformed Teradata scripts and stored procedures to SQL and **Python running on Snowflake's** cloud platform.
- Analyzed the system requirement specifications and in client interaction during requirements specifications.
- 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 Dynamo DB.
- Providing daily reports to the Development Manager and participate in both the design phase and the development phase. Utilized Agile Methodology and SCRUM Process.

Environment: AWS, Hadoop, Python, Pyspark, SQL, Snowflake, Data bricks/Delta Lake, AWS S3, AWS Athena and AWS EMR.

Carbon Health (San Francisco, CA) **Apr 2018 - Apr 2019**
Role: Data Engineer

Responsibilities:

- Responsible for building scalable distributed data solutions using **Hadoop**.
- Involved in Agile Development process (Scrum and Sprint planning).
- Handled **Hadoop cluster** installations in Windows environment.
- Migrated data warehouses to **Snowflake Data warehouse**.
- Defined virtual warehouse sizing for **Snowflake** for different type of workloads.
- Extracted data from data lakes, EDW to relational databases for analyzing and getting more meaningful insights using **SQL** Queries and **PySpark**.

- Designed, developed and did maintenance of data integration programs in a **Hadoop** and RDBMS environment with both traditional and non-traditional source systems.
- Developed MapReduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
- Wrote Sqoop Scripts for importing and exporting data from RDBMS to HDFS.
- Wrote **Python** scripts to parse XML documents and load the data in database.
- Written **DDL** and **DML** statements for creating, altering tables and converting characters into numeric values. Performed data cleaning and data manipulation activities using NOSQL utility.
- Worked on Data load using **Azure Data factory** using external table approach.
- Automated recurring reports using **SQL** and **Python**.
- Developed scripts in **Big Query** and connecting it to reporting tools.
- Designed workflows using **Airflow** to automate the services developed for Change data capture.
- Carried out data transformation and cleansing using **SQL** queries and **PySpark**.
- Used **Kafka** and **Spark** streaming to ingest real time or near real time data in HDFS.
- Worked related to downloading Big Query data into Spark data frames for advanced **ETL** capabilities.
- Participated in daily stand-ups, bi-weekly scrums and PI panning.

Environment: Hadoop 3.3, Big Query, Big Table, Spark 3.0, Sqoop 1.4.7, ETL, HDFS, Snowflake DW, Oracle SQL, MapReduce, Kafka 2.8 and Agile process

Fortinet (Sunnyvale, CA) **Jan 2017 - Mar 2018**
Role: Data Engineer

Responsibilities:

- Designed and deployed scalable, highly available, and fault tolerant systems on **Azure**.
- Led the estimation, review the estimates, identify the complexities and communicate to all the stakeholders.
- Involved in complete SDLC life cycle of big data project that includes requirement **analysis**, **design**, **coding**, **testing** and **production**.
- Defined the business objectives comprehensively through discussions with business stakeholders, functional analysts and participating in requirement collection sessions.
- Implemented end-to-end systems for **Data Analytics**, **Data Automation** and integrated with custom visualization tools. Migrated on-primes environment on Cloud using **MS Azure**.
- Designed the business requirement collection approach based on the project scope and SDLC (Agile) methodology. Moved data to **Azure Data Lake** to **Azure data warehouse** using PolyBase.
- Created external tables in ADW with 4 compute nodes and scheduled.
- Extensively used Agile Method for daily scrum to discuss the project related information.
- Worked with data ingestions from multiple sources into the **Azure SQL data warehouse**.
- Transformed and loading data into **Azure SQL Database**.
- Configured Spark streaming to receive real time data from the **Kafka** and store the stream data to HDFS.
- Development and maintenance of **data pipeline** on Azure Analytics platform using **Azure Databricks**.
- Developed a **data pipeline** using **Kafka** to store data into **HDFS**.
- Implemented Kafka producers create custom partitions, configured brokers and implemented High level consumers to implement data platform. Created Airflow Scheduling scripts in **Python**.
- Maintained NoSQL database to handle unstructured data, clean the data by removing invalidate data, unifying the format and rearranging the structure and load for following steps.
- Developed purging scripts and routines to purge data on **Azure SQL Server** and **Azure Blob storage**.
- Resolved the data type inconsistencies between the source systems and the target system using the Mapping Documents.
- Maintained data storage in **Azure Data Lake**.
- Written and executed customized **SQL** code for ad hoc reporting duties and used other tools for routine report generation.

Environment: Spark 2.8, Kafka 2.6.2, Apache Airflow 1.10, Azure SQL DB, Azure DW, Azure Data Lake, Azure Data factory, Python 3.7, XML, Azure Databricks, T-SQL and Agile process.

Edward Jones (St. Louis, MO) **May 2016 - Dec 2017**
Role: Data Engineer

Responsibilities:

- Collaborated with data architects to understand and get the data requirements for the project model as per the business.
- Developed Python script to hit Rabbit MQ API and extract data in JSON format and load into Spark RDD. Developed Spark program using PySpark, to handle Streaming data and load data Azure Events Hubs.
- Developed Talend ingestion frameworks to ingest between Teradata, MSSQL, HDFS, Azure Blobs and Azure Data warehouse.
- Extract Transform and Load data from Sources Systems to Azure Data Storage service using a combination of Azure Data Factory and ingest data Azure Blob storage and processing the data in Azure Databricks.
- Responsible for estimating the cluster size, monitoring and troubleshooting of the Spark Databricks cluster. Loaded data into Azure SQL database using Azure Databricks.
- Processed real time structured data using stream analytics in ADW in order to merge with the data that is being ingested from Hadoop to ADW.
- Ingested structured data from MySQL, SQL Server to HDFS as incremental import using Talend jobs. These imports are scheduled to run in a periodic manner.
- Responsible for creating Hive tables, loading the structured data resulted from Map Reduce jobs into the tables and writing hive queries to further analyze the logs to identify issues and behavioral patterns.
- Tuned the performance of Hive data analysis using clustering and partitioning of data with respective to date, location.
- Developed job workflows in CAWA to automate the tasks of loading the data into HDFS.
- Responsible to process and derive the data which is needed to build the customer defined metrics that are displayed in Power BI.

Environment: Python, HDFS, PySpark, Hive, Rabbit MQ, SQL, Azure (ADF, Blobs, SQL Data warehouse), MSSQL, Teradata, Power BI

Wissen Infotech (India) **July 2014 - Dec 2015**
Role: Data Analyst

Responsibilities

- Responsible for gathering data migration requirements. Identified problematic areas and conduct research to determine the best course of action to correct the data.
- Analyzed problem and solved issues with current and planned systems as they relate to the integration and management of order data. Involved in Data Mapping activities for the data warehouse.
- Analyzed reports of data duplicates or other errors to provide ongoing appropriate inter-departmental communication and monthly or daily data reports.
- Monitor for timely and accurate completion of select data elements.
- Collected, analyzed and interpreted complex data for reporting and/or performance trend analysis.
- Monitor data dictionary statistics.
- Involved in analyzing and adding new features of Oracle 10g like DBMS_SHEDULER, Create Directory, Data pump, CONNECT_BY_ROOT in existing Oracle 10g application.
- Archived the old data by converting them in to SAS data sets and flat files.
- Extensively used Erwin tool in Forward and reverse engineering, following the Corporate Standards in Naming Conventions, using Conformed dimensions whenever possible.
- Enhanced smooth transition from legacy to newer system, through change management process.

- Planned project activities for the team based on project timelines using Work Breakdown Structure.
- Compared data with original source documents and validate Data accuracy.
- Used reverse engineering to create Graphical Representation (E-R diagram) and to connect to existing database. Generated weekly and monthly asset inventory reports.
- Created Technical Design Documents, Unit Test Cases. Wrote SQL Scripts and PL/SQL Scripts to extract data from Database to meet business requirements and for Testing Purposes.
- Wrote complex SQL queries for validating the data against different kinds of reports generated by Business Objects XIR2.
- Involved in Test case/ data preparation, execution and verification of the test results.
- Created user guidance documentations. Created reconciliation report for validating migrated data.

Environment: UNIX, Shell Scripting, XML Files, XSD, XML, SAS, PL/SQL, Oracle 10g, Erwin 9.5, Autosys