Sujith Revilla
Data Engineer
revillasujith999@gmail.com

CONTACT: (737) 471-4704

PROFILE SUMMARY

- Around 5+ years of experience in Data Engineering, Data Pipeline Design, Development, and Implementation as a senior data Engineer.
- Proficient in data processing using Python, PySpark, Scala, and SQL, with a strong background in designing and implementing complex ETL pipelines.
- Experienced in managing big data technologies such as Spark RDDs, and Apache Kafka for efficient data ingestion and processing and Hadoop Distributions (Map R, Hortonworks, cloudera) to implement it's new features.
- Skilled in cloud platforms, including Azure, AWS, and GCP, focusing on designing scalable and cost-effective data solutions in the cloud.
- Experienced in Oracle Data Integration tools and technologies, responsible for designing, developing, and maintaining data pipelines and ETL processes to ensure efficient data integration and analysis.
- Expertise in ETL tools such as SSIS, Informatica, and Talend for data integration and transformation, ensuring data quality and consistency.
- Proficient in working with data warehouses like Snowflake, Redshift, and BigQuery, optimizing performance and data retrieval. Azure Data Factory (ADF), Integration Run Time (IR), Relational Data Ingestion, File System Data Ingestion.
- Experienced in define and implementing data models, tables, and views within Azure Synapse, ensuring optimal data storage and query performance.
- Worked closely with the data modeling team to create and maintain data models using DAX, enhancing the performance of Power BI reports.
- Design, develop, and maintain data pipelines to ensure efficient and scalable data processing for machine learning and AI applications.
- Proficient in data analysis using Pandas and NumPy, enabling data-driven decision-making and insights generation.
- Architecture, RDBMS and Data Modeling concepts, a specialist in SQL tuning and Performance.
- Experience working with Microsoft Azure Cloud services: Azure Data Lake storage Gen1 & Gen2, Azure Data Factory & other services.
- Skilled in using Databricks for data processing and ETL tasks.
- Familiar with data security, performance optimization, and collaboration using Databricks.
- Well experienced in understanding of Spark Architecture with Databricks, Structured, Streaming, settingup AWS and Microsoft Azure with Databricks.
- Proficient in data visualization tools like Tableau, Power BI, and Looker to create interactive and insightful data dashboards.
- Experienced in setting up logging and monitoring solutions using Splunk, ELK, and Azure Monitor for real-time data tracking and issue resolution.

TECHNICAL SKILLS

Data Processing: Python, PySpark, Scala, SQL, Impala, Hadoop, Spark RDDs, Apache Kafka

Cloud Platforms: Azure, AWS, GCP
ETL Tools: SSIS, Informatica, Talend

Data Warehouses and Modelling: Snowflake, Dimensional data modelling, Redshift, BigQuery

Data Formats: JSON, XML

Database Management: MySQL, MongoDB, SAS, Oracle, Cassandra, PostgreSQL, SQL Server, Cosmos DB

Data Analysis: Pandas, NumPy

Infrastructure as Code: AWS CloudFormation, Terraform, ARM templates

Containerization & Orchestration: Docker, Kubernetes

Data Visualization: Tableau, Power BI, Looker

Logging and Monitoring: Splunk, ELK, Azure Monitor

Version Control & CI/CD: Git, GitHub, Jenkins, Azure DevOps

Project Management: Jira, Agile, Scrum

WORK EXPERIENCE

Publicis Sapient,_Arlington,VA **Apr 2021 – Present**
Data Engineer
Roles & Responsibilities:

- Analyze, design, and build Modern data solutions using Azure PaaS service to support visualization data.
- Spearheaded the design and development of data pipelines using Python, PySpark, and Scala, ensuring efficient data extraction, transformation, and loading (ETL) processes.

- Managed complex data integration tasks by leveraging SQL, Hadoop, and Spark RDDs, optimizing data processing performance for large-scale datasets.
- Orchestrated data workflows in Azure Data Factory, Azure Databricks, and Apache NiFi, guaranteeing seamless data movement and transformation across cloud platforms and Hadoop clusters using Cloudera Manager.
- Implemented real-time data streaming solutions using Apache Kafka, processing high-velocity data streams for immediate insights.
- Contributed to Snowflake data warehousing tasks, including data loading and querying, and used PostgreSQL, ensuring data availability and accuracy.
- Designed and implemented data pipelines to extract, transform, and load (ETL) large datasets from various sources into a centralized data warehouse.
- Developed and optimized DAX formulas to create complex calculations, key performance indicators (KPIs), and business metrics in Power BI for data visualization.
- Developed and fine-tuned Hive and Pig scripts for data transformation, enabling efficient query processing in a Hadoop ecosystem.
- Involved in designing the Data pipeline from end-to-end, to ingest data into the Data Lake.
- Assisted in designing & developing data lake and ETL using python and Hadoop ecosystem.
- Automated data ingestion from various sources into Azure Cosmos DB, guaranteeing a consistent and reliable data repository.
- Designed and implemented data pipelines to extract, transform, and load (ETL) large datasets from various sources into Azure Synapse Analytics, leveraging its data warehousing and big data capabilities.
- Developed and maintained data integration processes using Azure Data Factory to orchestrate data movement and transformations within the Synapse workspace.
- Monitored and optimized data pipelines using Azure Monitor, ensuring data quality and reliability while meeting SLAs.
- Collaborated with cross-functional teams in an Agile environment, using JIRA and Azure DevOps for project management and version control with Git.
- Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
- Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Sql Activity.
- Created and maintained Azure Resource Manager (ARM) templates for infrastructure as code, automating cloud resource provisioning.
- Developed and maintained ETL processes using SSIS, ensuring data accuracy and consistency across the organization.
- Created interactive data visualizations and reports using Tableau, providing actionable insights to stakeholders.
- Collaborated with data architects to ensure data security, compliance, and governance by industry best practices.

Environment: Python, SQL, Hadoop, Hive, ADF, DAX, PySpark, Scala, Synapse, PostgreSQL, Azure, Azure Data Factory, Azure Databricks, Synapse, CosmosDB, AWS, DataLake Azure Monitor, Snowflake, RDBMS,ETL, SSIS, Tableau, Docker, Git, Agile.

Grainger, Lake Forest,IL **Jan 2020 – Mar 2021**
Data Engineer
Roles & Responsibilities:

- Designed and executed intricate data pipelines via AWS Glue for efficient transformation and loading of extensive data from diverse sources.
- Implemented Snowflake for data warehousing and analytics, overseeing the design and management of scalable and high-performance data storage solutions.
- Automated ETL processes through maintenance and optimization of Glue jobs and crawlers, ensuring seamless data processing and analysis.
- Designed and developed efficient stored procedures in AWS Redshift to automate complex data transformations, aggregations, and calculations.
- Conducted performance tuning and optimization of existing AWS Redshift stored procedures, enhancing system responsiveness and reducing query latency.
- Utilized SQL scripting within AWS Redshift's stored procedures to optimize query performance and minimize data movement.
- Engineered Spark, Hive, Pig, Python, Impala, and HBase data pipelines for seamless customer data ingestion and processing.
- Developed and sustained web applications using Django and Flask, adhering to Model-View-Controller (MVC) architecture for scalability and maintainability.

- Orchestrated Amazon EC2 instances creation, troubleshooting, and health monitoring, alongside other AWS services for multi-tier application deployment.
- Provided Linux and Windows cloud instances support on AWS, configuring Elastic IP, Security Groups, and Virtual Private Cloud.
- Implemented and managed Puppet for automated deployments and contributed to Chef and Puppet-based deployment strategies.
- Created OpenShift namespaces for on-premises applications transitioning to the cloud in OpenShift Pass environment.
- Virtualized servers using Docker for testing and development environments, streamlining configuration through Docker containers.
- Managed Docker clusters, integrating them with Amazon AWS/EC2 and Google's Kubernetes.
- Developed Jenkins CI/CD pipeline jobs for end-to-end automation, overseeing artifact management in Nexus repository.
- Configured JIRA as a defect tracking system, implementing workflows and customizations to enhance bug/issue tracking.

Environment: AWS, Python, SSH, Shell Scripting, JSON, Docker, Kubernetes, Red Hat Enterprise Linux, Terraform.

IQVIA, Durham, NC **May 2018 – Dec 2019**
Data Engineer
Roles & Responsibilities:

- Involved in business Requirement gathering, business Analysis, Design and Development, testing and implementation of business rules.
- Designing and Developing Azure Data Factory (ADF) pipelines to extract the data from Relational sources like, Oracle, SQL Server, DB2 and non-relational sources like Flat files, JSON files, XML files, Shared folders etc.
- Created Pipelines in ADF using Linked Services/Datasets/Pipeline to perform ETL on data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
- 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.
- Creating pipelines, data flows and complex data transformations and manipulations using Azure Data Factory (ADF) and PySpark with Databricks.
- Experience in optimizing Delta Lake performance by using partitioning, indexing, and caching techniques, and knowledge of advanced Delta features such as Delta streaming and Delta commands.
- Develop Azure Databricks notebooks to apply business transformations and perform data cleansing operations.
- Develop Databricks Python notebooks to Join, filter, pre-aggregate, and process the files stored in Azure data lake storage.
- Developed data pulls from proposed architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.
- Responsible for estimating the cluster size, monitoring, troubleshooting of the Spark data bricks cluster.
- Developed data pulls from proposed architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.
- Used Azure Synapse for Data Analysis and Reporting.
- Created maintained SQL Agent Jobs for automating and scheduling ETL process and Database maintenance tasks.

Environment: Azure Data Factory, Azure Data Lake System, Azure Data Brics, Spark, PySpark, Hadoop, Hive, Sqoop, HDFS

EDUCATION DETAILS:

Master's in computer science, Texas state university