Pravalika S

Atlanta, GA | +1(816)2638324 | spravalika95@gmail.com | GitHub | LinkedIn

PROFESSIONAL SUMMARY

- Over 5+ years of experience in Data Engineering professional in building data pipelines, agile methodology, interpreting and analyzing trends in the data using trending techniques and building machine learning models.
- Knowledgeable in Azure Data Services, including Azure SQL Database, Azure Data Lake, and Azure Synapse Analytics, with a proven track record of designing and implementing data solutions on the Azure platform.
- Experienced in developing complex SQL applications, managing cloud platforms (GCP, AWS), and deploying on Linux-based environments, complemented by proficiency in service-oriented architectures, API design, CI/CD, and testing frameworks for seamless system integration and reliability.

TECHNICAL SKILLS

- Programming Languages: Python, R Language, Java, C, Pyspark, SQL, R, HTML, Scala, GoLang, ReactJS, javascript
- Competencies & tools: Jupyter, Databricks, Docker, RStudio, Git, PyCharm, Terraform, VS Code, PowerBI, Tableau
- Big Data Technologies: Spark, Hadoop, MapReduce, Spark-SQL, Hive, kafka, Pytorch, Pandas, NumPy
- Cloud Technologies: Azure Blob, Databricks, Azure Synapse, AWS S3, Redshift, EC2, Lambda
- Databases: MySQL, NoSQL (Document DB, Column family DB, Graph DB), MongoDB, Oracle
- Machine Learning: ANN, CNN, LLM models, Statistics (Hypothesis testing, ANOVA), Regression (Linear, Logistic, Ridge, Lasso), Decision Tree, Random Forest, XG Boost, K-Means, KNN, Support Vector Machine, A/B Testing, NLP, Scikit, Seaborn, Plotly

EXPERIENCE

MINDTREE *April 2022 – April 2023*

Data Engineer (Skills: Analysis, Power BI, GCP, AWS, Tableau, SQL, Kafka, Pyspark, , Azure Databricks, ETL/ELT, Frameworks, Informatica, O9 solution)

- Built and optimized scalable data processing workflows using Azure Databricks, by using Spark for data ingestion, transformation, and analysis., reducing migration time by 50%.
- Designed and implemented end-to-end data pipelines using Pyspark and Python on Azure Databricks, resulting in improved data processing efficiency by 30%.
- Developed and executed SQL queries, transformations to extract and transform data from multiple sources, resulting in a 20% increase in data accuracy and quality. Skilled in creating interactive Tableau dashboards and visualizations, connecting to diverse data sources, and customizing visualizations to meet specific business needs.
- Automated file transfer process by automating Python workflow to move files seamlessly to any designated location and trigger email notification, resulting in a 30% reduction in manual effort.
- Implemented comprehensive data extraction and transformation processes across diverse sources, achieving a 20% increase in data accuracy and quality.
- Developed custom-built ETL solution, batch processing, and real-time data ingestion pipeline to move data in and out of the Hadoop cluster using PySpark and Shell Scripting.
- Proficient in advanced SQL querying for efficient data analysis and manipulation from large-scale databases like Teradata, demonstrating a strong grasp of database concepts and data warehousing principles.
- Designed and deployed data pipelines using Azure cloud platform (HDInsight, Data Lake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH, and Data Storage Explorer) for data movement and transformations.
- Responsible for transitioning the data models from informatica to PySpark data models in Azure databricks. Applied statistical modeling and ML techniques for insightful data analysis.
- Automated file transfer process by automating Python workflow to move files seamlessly to any designated location and trigger email notification, resulting in a 30% reduction in manual effort.
- Optimized high-volume SQL queries and integrated Python libraries such as Pandas and NumPy for complex data manipulation and analysis resulting in a 40% reduction in execution times and a 35% boost in workflow efficiency.
- Proficient in utilizing the Microsoft ETL data ingestion framework, particularly SQL Server Integration Services (SSIS) DevOps, to efficiently manage and automate the extraction, transformation, and loading of data, ensuring streamlined data processing workflows and optimized performance.

GENPACT *Jan 2020 – Apr 2022*

Data Analyst (Skills: Analysis, Agile Methodologies, Power BI, GCP, AWS, Tableau, SQL and Kafka)

- Proficient in utilizing the Microsoft ETL data ingestion framework, particularly SQL Server Integration Services (SSIS) DevOps, to efficiently manage and automate the extraction, transformation, and loading of data, ensuring streamlined data processing workflows and optimized performance.

- Adept at using Microsoft SQL Server Reporting Services (SSRS) and SQL Server Analysis Services (SSAS) to design, develop, and deploy complex business intelligence solutions that improve data visualisation and help stake holders make well informed decisions.
- Collaborated with with cross-functional teams to develop and execute a data migration plan, successfully moving data from source staging to staging, and then to enterprise and production systems and created data models for data flow.
- Conducted thorough system analysis and design to identify technical solutions aligned with business requirements, resulting in a 15% increase in system efficiency.
- Utilized SQL for data analysis and database management, extracting insights and providing data-driven recommendations to improve business processes.
- Managed projects following Agile methodologies, coordinating with development teams and stakeholders to deliver high-quality solutions on time and within budget.
- Proficient in developing complex SQL applications, managing cloud platforms (GCP, AWS), and deploying on Linuxbased environments. Proficient in using advanced Excel functions and features for data analysis and reporting, including data cleaning, manipulation, and transformation.
- Skilled in creating complex Excel models and performing statistical analysis. Expertise in optimizing SQL queries for enhanced performance and efficiency.
- Experience with Python/R for data analysis, visualization, and modeling, demonstrating a strong foundation in statistical analysis techniques and machine learning algorithms. Proven track record of building and interpreting data models to support informed business decision-making.

HCL *May 2018 – Jan 2020*

Data Engineer (Skills: Azure, Kafka, Aws, Pyspark, Python, Machine Learning, ETL pipelines, Git, REST API)

- Developed a model to predict customer Wishlist for shopping using supervised Machine Learning, using historic data of startups, performed data cleaning, feature engineering, implement a web-based client-side.
- Leveraged AWS (Amazon Redshift, Amazon S3, AWS Glue) for scalable data architectures, storage, processing, and analysis.
- Automated data workflows with AWS Step Functions, Lambda, and Batch. Managed complex software applications with Jira and Git.
- Monitored system health and data processes using AWS CloudWatch. Created high-performance RESTful API endpoints for easy access to Sagemaker models, ensuring availability and data protection. Revamped testing infrastructure, reducing test time by 50% and achieving over 90% coverage. Streamlined data cleaning and preprocessing for NLP science models.
- Leveraged AWS (Amazon Redshift, Amazon S3, AWS Glue) for scalable data architectures, storage, processing, and analysis.
- Conducted thorough system analysis and design to identify technical solutions aligned with business requirements, resulting in a 15% increase in system efficiency.
- Carried out A/B Testing to improve the Click-through rate and effectiveness of marketing campaigns; reduced marketing budget by ~ 30%.
- Managed complex software applications with Jira and Git. Monitored system health and data processes using AWS CloudWatch.
- Collaborated with data scientists to manage and store data masking it from loss and theft by implementing data encryption and also how to organize data and give the resultant data to them for analysis and training machine learning models
- Completed 15+ CI/CD pipelines as per client requirements. Engineered a Model-based Collaborative Filtering System to recommend our product to customers which increased cart value by 12%; helped ameliorate average session time from 1.47 to 3.18 minutes.

PROJECTS

Real-Time Data Pipeline for E-commerce Platform

Developed a real-time data pipeline for an e-commerce platform to analyze user behavior, manage inventory, and optimize marketing strategies. Used Kafka to collect real-time user events such as page views, clicks, and purchases.Used Python and PySpark to preprocess and clean the streaming data, performed ETL (Extract, Transform, Load) operations, and aggregate the data. Stored the processed data in AWS S3 and maintain data catalog using AWS Glue. Loaded the transformed data into Amazon Redshift for analytical queries and reporting. Created dashboards in PowerBI/Tableau to visualize key metrics such as sales, user engagement, and inventory levels. Implemented monitoring and alerting using AWS CloudWatch. Deployed the pipeline using AWS Lambda functions and schedule regular updates. Continuously monitored and optimized the pipeline for performance and cost efficiency.

Data Migration and Transformation for Financial Institution

Migrated and transformed legacy data from on-premises databases to cloud-based data warehouses for a financial institution, enabling advanced analytics and reporting capabilities. Analyzed existing data sources and defined migration strategy considering data volume, complexity, and business requirements. Used Azure Data Factory to extract data from on-premises databases and transformed it into a suitable format for migration. Transferred the transformed data to Azure SQL Database and Blob Storage for staging and validation. Developed Python scripts to further clean and transform the data according to target schema and business logic. Loaded the processed data into AWS Snowflake data warehouse for analytics and reporting. Performed extensive validation using MS-Excel and SQL queries to ensure data accuracy and integrity. Created PowerBI dashboards to visualize financial metrics, trends, and KPIs for stakeholders and decision-makers. Deployed the migrated data solution on cloud infrastructure and established monitoring and maintenance procedures to ensure ongoing data accuracy and performance.

Building Scalable Data Lake for Healthcare Analytics

Designed and implemented a scalable data lake solution for a healthcare organization to centralize and analyze diverse data sources for operational insights and decision-making. Used Apache Kafka to ingest real-time data streams from various healthcare systems such as Electronic Health Records (EHR), IoT devices, and patient monitoring systems. Designed a scalable data lake architecture using AWS S3 to store structured, semi-structured, and unstructured data in its native format. Utilized AWS Glue for automatic discovery, cataloging, and organization of data assets within the data lake. Developed Python scripts and PySpark jobs to cleanse, transform, and enrich the raw data before storing it in the data lake. Enabled ad-hoc querying and analytics on the data lake using AWS Athena and Redshift Spectrum, allowing analysts to derive insights from large datasets without the need for data movement. Implemented machine learning models on DataBricks to perform predictive analytics for patient outcomes, disease detection, and healthcare resource optimization. Created interactive dashboards in Tableau to visualize key performance indicators (KPIs), patient demographics, disease trends, and healthcare utilization metrics. Established data governance policies, access controls, and encryption mechanisms to ensure compliance with healthcare regulations such as HIPAA. Deployed the data lake solution on AWS infrastructure and set up monitoring and alerting using AWS CloudWatch for proactive issue resolution and performance optimization.

EDUCATION

University Of Central Missuori

Master of Science in Computer Science

SR College of Engineering

Bachelor of Science in Computer Science and Engineering

AWARDS AND ACCOMPLISHMENTS

- Best Innovative Employee of the Data Engineering team in a year at Genpact.
- Event Lead (Annual College Fest), Organized two intra-college coding competitions with footfall of 200+ participants.
- Part of Sulakshya Seva Samithi NGO (3+ years), aimed at solving hunger problems and COVID-19 vaccination awareness.
- Runner-Up (2nd/74 students), Accenture Code-a-Thon.