CCA cloudera
SPARK & HADOOP DEVELOPER

RAVI TEJA VORUGANTI

Lead AI Data Engineer | Gen AI & RAG architect | Building Scalable, AI powered Enterprise Data Platform

Summary

Overall, 9+ years of experience in Data Engineering working on the various technologies for data engineering solutions along harnessing latest GenAI agents on cloud platforms and on premises. Associated with Excelra for the past 6 years, contributing to multiple projects to GSK/other clients and internal assignments. Was among the first team to set up the Hadoop and Spark cluster to conduct POC's on the big data infrastructure at Excelra.

Skills

- Data Engineering experience in building Data intensive applications with Medallion and Data Mesh architecture often implementing ETL with all the necessary Data acquisition, data ingestion, data cleaning, enrichment & transformation of the data to build data platforms with high performance and scalability.
- Databricks Pyspark, MLflow with GenAI, Hadoop Big data ecosystem tools like Hive, Sqoop, flume, NiFi, HBase, Streaming data solutions with Apache, Kafka, Confluent Kafka and Strimzi Kafka on Kubernetes. Management and deployment of workloads on Kubernetes and its management. Azure cloud services like Function, HD Insights, Synapse analytics, Stream Analytics Jobs, EventHub, Logic Apps, Azure DevOps CI CD with GitHub, Denodo Virtualization, Apache Airflow, Astronomer Airflow, CDAP, Snowflakes Datawarehouse, Data Modelling.
- Hands-on experience in Python, Scala & R languages.
- Experience in the AWS and GCP cloud services.
- Project Management adhering to Agile methodology using Jira and confluence.

Experience

- Working as a Lead Data and AI Engineer with Excelra BV, London, UK from 01st Sept 2023 to Present
- Worked as Senior Azure Data Engineer with Excelra Knowledge Solutions from 19th July 2019 to 31st August 2023
- Worked as Application Development Consultant with NTT Data Services from 10th Sept 2016 to 16th July 2019

Professional Experience

Project: Eli Lilly – Strategic Capabilities & Data & Analytics
Nov 2025 – Present
DSAER AWS comments App Development
Environment & Tools: AWS Lambda Function, AWS API gateway services, PostgreSQL, AWS Aurora DB, DevOps with GitHub Actions, Serverless Framework, Python, Node.js

Roles & responsibilities:

- Part of the DIVA AWS Project, a requirement to integrate the DSAER module (DIVA Serious Adverse Events Report) with the comments functionality that enables Data Managers to put the comments for several studies dynamically and track their progress.
- Development of a new Data API over the existing pipelines using AWS Lambda functions and Aurora PostgreSQL. The Power Automate Custom connector is also developed to connect our Data API and support the integration.
- New Feature enablement to the DSAER module that will help Data managers to automatically track and send the comments for each study and the proactively mitigate the risks associated with ongoing studies.

Project: Eli Lilly – Strategic Capabilities & Data & Analytics
Aug 2025 – Oct 2025
SMILe (Study Management Insights & Learning) - Study& Site Prediction:
Environment & Tools: Power BI, Microsoft Data Fabric, Power Automate, Python (pandas), VS Studio Code
Roles & responsibilities:

- Transitioning manual Excel-based R outputs into a relational schema with minimal disruption to Power BI dashboards—enabling scale, performance, and ML-readiness.
- We have implemented a modular consolidation design, reducing 54 tables to 28 using a dynamic mapping sheet and Python automation. This ensures schema flexibility and seamless AWS integration.
- Achieved backend compatibility with existing dashboards, avoiding major overhaul in view of the timelines.
- Modernizing the backend data architecture that minimizes the latency and streamline data workflows.

Project: Eli Lilly – Strategic Capabilities & Data & Analytics
Feb 2025 – Present
AI powered Program Automation Tool for Stat Analysts
Environment & Tools: Azure AI Search, Azure Open AI, Azure, Langchain, FAISS Vector search, Knowledge store, Python, Streamlit UI, Posit Rstudio
Roles & responsibilities:

- Lead the team to be able to generate accurate and reliable R programs for the SDTM/ADaM variables of a clinical study that saves hundreds of Human hours for Statistical Analysts.
- To overcome the problem, we developed an AI-powered ADaM Program Automation Tool that uses Hybrid Retrieval-Augmented Generation (RAG), vector databases, and large language models (LLMs) to generate high-quality R code directly from a specification sheet. By combining FAISS-based semantic search, Azure OpenAI embeddings, and a feedback-driven learning loop, we created a solution that reduces manual effort, improves consistency, and builds a reusable knowledge base over time.
- Development of an AI powered ADaM Program Automation Tool that can automatically generate the R programs accurately from an Input Specification sheet and feed the output back into the vector database that can enrich the knowledge base further.
- Saving around 40% of efforts improving the productivity and saving inference costs of LLM's in long run.
- Saved ~70% manual programming effort, significantly accelerating data pipeline delivery.

Project: GSK Usage Product
Sept 2021 – Feb 2025
Azure GenAI transformation for Behaviour Analytics
Environment & Tools: Azure AI Search, Azure Translator, Azure Databricks, Vector search, Knowledge store
Roles & responsibilities:

- Transformed the traditional NLP ML model for text categorization by using Vector Search via Azure AI, indexing knowledge stores and integrating RAG functions. I'm now focused on productionizing the workflow using Azure APIM and Azure Kubernetes Service.
- An innovative approach to modernizing the traditional solution into a more advanced Gen AI approach that resulted in a significant increase in the accuracy of the model output and was more intuitive.
- Achieved 70% savings from the efforts of retraining and development.

Marketplace Domain Registration & Management Metrics
Environment & Tools: Databricks, Snowflakes, Log Analytic workspace, Azure Data Factory, ADLS v2, Azure DevOps
Roles & responsibilities:

- Built the metrics dashboard based on ETL data modelling solution for the domain owners and leadership to identify the missing key metadata fields led to the discovery of the data domains that improved the Marketplace value by 60%.
- The usage of the datasets doubled by the greater visibility that accelerated the research through better funding and promoted the value of the data domains.

Denodo structured usage analytics
Environment & Tools: Databricks, Snowflakes, Log Analytic workspace, Azure Data Factory, ADLS v2, Azure DevOps
Roles & responsibilities:

- Pioneered the cloud migration initiative at GSK by deploying a new cloud data engineering solution on Azure, utilizing various Azure services, as well as Azure Databricks, Snowflake, and Denodo virtualization platforms.
- Employing watermarking techniques, I ingested a terabyte of data from log analytics workspace, addressing Log Analytics fetch size limits. The data was then stored in Snowflake, providing a centralized platform for analytics tools and achieving annual savings of around $200,000.
- Virtualized the data using the Denodo interface, allowing for programmable consumption and adherence to the FAIR principles.

Marketplace Search Usage Text classification to topic streams
Environment & Tools: Python, NLP, Databricks MLflow, Databricks serving endpoint
Roles & responsibilities:

- Implemented text classification model on the Databricks MLflow platform to categorize user search text into relevant study topics, enabling stakeholders to identify trending subjects and prioritize requirements that need immediate attention.
- Deployed the model on databricks serving endpoint to make the model available for the end users as API that improved the productivity by 50%.

OMP Data pipeline refactoring
Environment & Tools: Azure Databricks, Snowflake cloud, ADF, Log Analytics workspace
Roles & responsibilities:

- Migrated the existing data pipeline into a new architecture guided by the Architecture Review Board to adhere to the stratsegic roadmap.

• Reduced the effort by 50% to maintain the event hubs and automated the ingestion process.

ADLS v2 file-based usage analytics
Environment & Tools: Databricks, Snowflakes, Log Analytic workspace, Azure Data Factory, ADLS v2, Azure DevOps
Roles & responsibilities:

• Designed the data lake over Snowflake to parse the usage logs for establishing risk management and building insights
on the usage of file-based data products.

• Attained the milestone of establishing the monitoring alerting process to protect the study data from unauthorized
access and promote data-driven insights on the data product usage.

Project: Data migration to Azure cloud Platform
Duration: Feb'23 – Jun'23
Client: GSK Digital & Tech
Environment: Cloudera Hadoop Data Platform, Strimzi Kafka 0.18.0(Apache Kafka distribution) on Rancher Kubernetes env, Databricks Pyspark and Delta Lake tables, ADLS V2 storage, Azure Eventhub
Tools/Technology: Kafka 2.6, Python, Elastic Cluster, Kibana Dashboard, Logstash with Kafka for Usage log ingestion.
Description:
GSK Journey to cloud direction to modernize and migrate to the Code orange Platform with Databricks pyspark and Azure cloud services.
Roles & Responsibilities:
Led the team in designing the architecture of the different components of the platform able to serve the purpose of
building the modern cloud data analytics platform.
• Drive the team to migrate the on-premises Hadoop-based platform to the Azure cloud.

Project: GSK Behavior Analytics Platform
Duration: Oct'22 – Feb'23
Client: GSK R&D Tech
Environment: RDIP Data Platform, Strimzi Kafka 0.18.0(Apache Kafka distribution) on Rancher Kubernetes env, ELK tech for visualization & report generation.
Tools/Technology: Kafka 2.6, Python, Elastic Cluster, Kibana Dashboard, Logstash with Kafka for Usage log ingestion
Description:
With the profound experience in the usage and data insights built for the leadership team and business stakeholders we are currently working on integrating the usage metrics of different applications over GSK hybrid cloud into ELK platform for enhanced dashboards, greater grains of data visibility, better visualization and provide prescriptive analytics along with the informative dashboards.
Roles & Responsibilities:

- Have been playing a lead role in defining and architecting the different components of the platform to be able to serve the purpose of building the prescriptive analytics platform.
- Lead the team to deliver a POC with one of the GSK usage dashboard to be completely migrate to ELK and build visualizations on Kibana with Vega plugins
- Cutting edge alerting and monitoring systems to predict anomalous behavior and detect breach for the respective data usage using machine learning models
- Designing and implementing the enhanced insights for the product teams to be more powerful and enable them to take data driven decisions.

Project: GSK Marketplace-Usage Product Analytics
Duration: Jun'20 – Sept'22
Client: GSK R&D Tech
Environment: RDIP Data Platform powered by Cloudera Hadoop HDFS, Strimzi Kafka 0.18.0(Apache Kafka distribution) on Rancher Kubernetes env, PowerBI for visualization & report generation, ELK system integration
Tools/Technology: Hive/Impala, Hadoop HDFS, Hue, Kafka 2.6, CDH 6.3.3, Streamsets Data Collector, Airflow DAG's, Python, ELK stack
Description:
R&D Data Platform Plus Usage analyzes usage data and metadata from multiple source systems, performs data analytics and transforms the data as per the business requirement. The raw usage data captured and published by asset owners/product teams are stored in a usage repository in HDFS initially, and then transformed and stored in a transformed DB in RDDP. The transformed DB is in the consumption layer and connects to the Power BI. Later integrated with Azure services to transform to RDIP+ usage.
Roles & Responsibilities:

- Connect to different types of external data sources (structured & unstructured) to fetch data using Kafka connect connectors to Kafka topics in near real-time.
- Deploy and manage the Strimzi Kafka on Kubernetes cluster using operators provided by Strimzi.
- Perform data conversions to Avro and Parquet files using Confluent schema registry based on requirement for data processing or data repository.
- Develop airflow DAG and deploy to schedule daily data load jobs using ssh & bash operator.
- Handling the incoming data effectively using the data partitioning and offset management.
- Ingest the data from the Kafka topics to Kerberos authenticated HDFS using HDFS sink connector and Streamsets pipeline.
- Load raw data from ingestion DB to Hive tables to analyze the data.
- Develop python scripts and sql queries to perform several transformations & aggregations on the hive tables and save resultant metrics to impala DB's that will source the data to PowerBI for various Dashboard visualizations and report generations.
- Responsible for deployment of the product into production from Github CI/CD pipeline and schedule jobs to run scripts to refresh the data in impala tables.

Project: Clinical Blobby
Duration: Dec'19 – May'2020
Client: Confidential
Environment: ADLS Gen2, Azure FileShare, ADF, Azure Functions, Azure KeyVault, AADDS & AD DS, Python, Postgres
Tools/Technology: ADLS Gen2, Azure FileShare, ADF, Azure Functions, Azure KeyVault, AADDS, AD DS, Python, Postgres
Description:
Clinical Blobby application is designed to provide a solution that can allow end users to create a new study, assign study to statisticians, download and upload studies and archival of completed studies. This solution should provide with the ability to complete study to experts PR/Monitors, Data Entry, Statisticians and Data Managers using studies from ADLS Gen2 to GitHub and to local system and vice versa. Mount Azure Files on user's desktops and manage ACL's.
Roles & Responsibilities:

- Load study Files into Azure FileShare
- Extract Metadata from the Study files using Azure Functions and load into Postgres tables
- Enabling Azure Fileshare mount on User Desktop using SMB
- Configuration of Azure Active Directory Domain Services (AADDS) to enable users to mount Azure FileShare.

- Implementation of PowerShell scripts to manage ACL’s on FileShare
- Automating the workflow using Azure Data Factory (ADF)

Company: Excelra Knowledge Solutions
Project: Mobius Data Ingestion
Duration: Oct’19 – Dec’19
Client: Confidential
Environment: Hadoop HDFS, PostgreSQL, Confluent Kafka
Description: Mobius data from the landing zone on HDFS is streamed to kafka topics and consumed by Kafka consumer into PostgreSQL Tables using the KAFKA - Connect plugin provided by Confluent Kafka (Open-Source distribution of the Apache Kafka).
Role: Created the topics to stream the data into it and kafka connect consumers to consume the data stored in the topics into PostgresSQL tables.

Project: GOSTAR, GOBIOM, IOCTA (Immuno Oncology Clinical Trial Analytics)
Client: Internal
Environment: AWS S3, Databricks Spark, Python, SparQL, RDF, GraphQL, NoSQL (MongoDB), OWL, Protégé
Description: GOSTAR is the quintessential Structure Activity Relationship (SAR) intelligence platform for drug discovery. From target profiling to hit identification and lead optimization, GOSTAR is the perfect resource for the medicinal and computational chemists capturing granular assay data across chemical, biological, pharmacological, and therapeutic dimensions. It also includes custom curation support for bespoke client needs, data preparation for AI/ML modeling and secure curation on the cloud of proprietary client data.
GOBIOM Biomarker Database is a comprehensive biomarker database that provides information on proteomic, Genomic, Biochemical, Imaging, Metabolite, Clinical Scoring scales and Cellular biomarkers for 18 different therapeutic area
Role:

- Data Ingestion from various Internal/External sources into AWS S3.
- Extraction of data points from the files and perform Semantic Data Conversions including identification of Things and Concepts in dataset using Databricks pyspark.
- Identifying Master Data Reference data attributes.
- Choosing URI Scheme for things and assigning column names/values in dataset using RDF.
- Load the URI enhanced data in MDM Querying the results using SparQL.
- Automating the above steps using a Pipeline.
- Experience with Semantic Web, including RDF/OWL and SPARQL

Company: Excelra Knowledge Solutions
Project: Gviz Evidence Package
Duration: Aug’19 – Dec’19
Client: GSK
Environment: Hadoop HDFS, PostgreSQL, Confluent Kafka
Description: Coordinate data points of Gviz graphs are stored in PostgreSQL table columns that generate
graphs. That column data will be loaded to the Kafka Topics using the KAFKA
Connect plugin. Feature provided by Confluent Platform (Open-Source distribution of the Apache Kafka)
Role: Using Kafka Connect Plugin we ingested the PostgreSQL column data into a new Topic then data is pushed into Hadoop HDFS as an evidence pack that will be stored as an archive of GViz graphs.

CCA
cloudera
SPARK & HADOOP
DEVELOPER

Achievements

- Distinction in the Paper Presentations held at the Envision 2016
- Received multiple client recognitions consistently over 3 years for driving innovation, leading cloud migrations and implementing streaming ingestion pipelines using Kafka on Kubernetes.

Certifications

- Azure Fundamentals | Databricks Lakehouse Fundamentals | Databricks AI Fundamentals | Cloudera CCA -175

Education

- MBA in Finance from BITS Pilani, India | GPA - 7.8/10
- Bachelor of Technology (B.Tech) in Computer Science Engineering from JNTU Hyderabad, India | GPA - 8.2/10