

# BHASKAR DAS

Worked at Business Intelligence Unit (BIU) of Axis Bank HQ, Worli, INDIA

## [ MY PROFILE / ROLES ]

Lead Big Data Engineer / Developer (Full Stack) | Data Scientist |  
| Senior Data Architect | Senior Analyst | AWS/Azure Cloud Architect/Management |  
| Lead Data Analyst | Business Intelligence | DevOps CI/CD

### CONTACT

Call / WhatsApp → +91 70 011 022 73 / +91 892 77 888 00  
Email → [TheBhaskar@Outlook.com](mailto:TheBhaskar@Outlook.com)  
Available to Join → Immediately !

### ACADEMIC BACKGROUND

MBA in Entrepreneurship → JNTU Hyderabad – First Class Distinction  
Bachelor of Engineering (B.E.) in ECE → UIT BU – T CPA 7.8 of Marks  
Diploma Engineering in ETCE → BIT – 72.5% of Marks

### GITHUB

GitHub – [www.github.com/TheBhaskarDas](https://www.github.com/TheBhaskarDas)

### PROFILE SUMMARY

As a **dynamic Techno-Functional IT Professional in executive management position** with 10 years of extensive experience in the IT industry, I possess advanced expertise across **various functionalities and technologies**. My proficiency spans **Data Analytics, Full Stack Data Engineering/Developer of Hadoop and Spark Ecosystems** with Core **DevOps** environment, employing **SQL, Python, PySpark and Scala**, delved into **Machine Learning Algorithms, Data Science, Computer Vision Engineering and Cloud Product Manager** by working with the Cross-functional team to create a roadmap, define features, and ensure the company is delivering the best experience possible. I have had always aligned with the **latest technological advancements**. I aspire to establish myself as a symbol of trust and reliability in the corporate arena. Throughout my career, I have successfully led numerous high-impact projects of Core **Banking, Insurance, Telecom, Healthcare/Health-Tech, Pharmaceutical, Retail Lending, Sales, Transport and IoT** sector, concurrently managed projects in **UAT and production spanning** resulting in increased efficiency and revenue growth.

### TECHNICAL EXPERTISE (Full Stack Architect, Full Stack Data Engineer, Data Science)

- **Big Data Ecosystems/Distributed Technologies** → Apache Spark v3, Spark Structured Streaming, PySpark, Kafka, Flafka, Hadoop v1 & v2, YARN, HDFS, MapReduce, HBase, Zookeeper, Hive, Sqoop, Oozie, Flume, ETL, Datawarehouse/DWH
- **Cloud Computing Services** → **Amazon Web Services** (S3, Athena, Glue, Custom Classifier, Quicksight, EMR, EC2, Redshift, Amazon Kinesis Data Firehose, Kinesis Data Streams, AWS Lambda, DynamoDB, DocumentDB, DB Migration Service, Boto3), **Microsoft Azure** (ADF, USQL, Databricks, HDInsight).
- **Statically/Dynamically Type Programming Languages** → Scala, Java, C, C++, Python
- **Dataset Involvement** → XML, JSON, Parquet, ORC, Avro, OpenCV, CSV, PDF etc.
- **Algorithms Involvement** → YOLOV8, KNN, Linear Regression, Logistic Regression, Decision Tree, Artificial Neural Network, Reinforcement Learning, Random Forest
- **Python Libraries / Machine Learning Frameworks** → TensorFlow, OpenCV, Ultralytics's YOLO v8, EasyOCR, SciPy, Pandas, NumPy, Boto3, Keras
- **Job Scheduler** → Airflow, Oozie
- **Visualization Tool** → AWS Quicksight, Power BI, Apache Hue, Kibana
- **Databases/Query Language** → NoSQL, Oracle 11g, SQL Server, HQL(Hive)
- **Tools** → SBT, Maven, MQTT, Confluent, Jersey REST, gson, JDOM2
- **IDEs** → Dbeaver, Zepplin, Eclipse, NetBeans, MS Visual Studio, IntelliJ IDEA, PyCharm, Jupyter Notebook in Google Colab
- **Platforms** → Windows, Unix, Linux- Ubuntu 16.x, CDH
- **DevOps Pipeline Deployment, Version Control Code Management, Workspace & Production Release Process Execution** → CI/CD, Git, Bitbucket, Jenkins, JIRA, Confluence. JSM, JPD
- **Agile Methodology Framework (Core + Technical)** → Agile Project Management, Scrum & Kanban Framework, Sprint Planning Execution, Backlog Grooming, User Story Mapping, Continuous Integration/Continuous Deployment CICD, Timeboxing, Retrospective Facilitation, Test-Driven Development (TDD), Behavior-Driven Development (BDD).
- **Leadership Experience** → Managing Cross-functional Team Across Multiple Locations, Stakeholder Engagement, Direct Client Engagement, Manageing SDLC (involved in the full life cycle i.e. scoping, designing, implementing, testing, deploying, and maintaining software systems across our products), Manage Firm Regulatory and Reputational Risk, Develop and Manage Talent Within the Team and High Quality Deliverables, SWOT Analysis, Key Metrics to Track Implementation Effectiveness, Launching Innovative and Go-To-Market Product Strategy, Product Development, Customer Feedback Focused Improvements and Delivers a High-Quality Experience. Also involved in Agile Project Management, Daily Scrum & Kanban Agenda, Sprint Planning, Reviews & Execution, Backlog Grooming, User Story Mapping, Continuous Integration/Continuous Deployment CICD, Timeboxing, Retrospective Facilitation, Conflict Resolution.

### PROJECT - 1 DESCRIPTION

Lead Data Engineer / Lead Data Architect  
Axis Bank - Business Intelligence Unit (BIU)  
Retail Lending | Database Lending | Digital Lending | Risk Assesment & Mitigation Strategies



**Core-Banking Domain Client:**  
Axis Bank, 3rd largest private sector bank in India offering entire spectrum of financial services for personal & corporate banking.

### MY ROLES & RESPONSIBILITIES:

- As a seasoned Lead Data Engineer at Axis Bank, I played a pivotal role in driving data-driven initiatives within the Retail Lending, Database Lending, Digital Lending, and Risk Assessment & Mitigation Strategies domains..
- **Data Engineering Leadership:** Led a team of data engineers in designing, developing, and maintaining robust data pipelines using cutting-edge technologies like Spark/PySpark, Kafka, Hadoop, Hive, Sqoop, and Hue.
- **Big Data Architecture:** Architected and implemented scalable big data solutions, including data lakes and data warehouses, to support critical business decisions.
- **Project Management, Innovation & Migration:** I manage end-to-end project execution, from monthly product release planning to cloud migration initiatives, my leadership fosters a zero-workaround environment, driving innovation across product development and release processes.
- **Data Governance and Security:** Ensured data quality, integrity, and security by implementing industry best practices and compliance standards.
- **Cluster & Performance Optimization:** Optimized data pipelines and query performance through code refactoring, indexing, and **cluster configuration** for monthly release processes, guaranteeing high performance and stability.
- **DevOps and Automation:** Implemented CI/CD pipelines using tools like Airflow, BitBucket, and Jenkins to automate data workflows and reduce manual intervention.
- **Collaboration, Compliance and Team Management:** Utilizing JIRA, Confluence, and BitBucket, Fostered a collaborative work environment, mentored team members, and effectively communicated with stakeholders at all levels. I uphold data security standards, enforcing best practices for data governance and regulatory compliance across all processes.
- By leveraging my expertise in data engineering, cloud technologies, and data governance, I consistently delivered high-quality data solutions that empowered the business to make informed decisions and achieve strategic objectives.

### PROJECT- 1 INFORMATION ( CORE BANKING DOMAIN )

- **Client & Payroll Company:** Axis Bank
  - **Employment Status:** Permanent
  - **Planned, Developed & Launched Functional Methodologies. Involved Tools, Platform and Technologies:**
    - **DevOps (JIRA, Confluence, BitBucket, Jenkins)**
    - **Used Jira Service Management (JSM) and Jira Product Discovery (JPD)**
    - Big Data Analytics
    - Python (Programming Language)
    - PySpark, Hadoop, Hive, Sqoop
    - Cloud Migration
    - Oracle & PL/SQL
    - Big Data Engineering
    - Memory Management
    - Troubleshooting
    - Performance Tuning
- In Summary, I bring a wealth of experience, strategic acumen, and technical mastery to drive excellence and innovation in the BIU.

### PROJECT - 2 DESCRIPTION

Policybazaar.com (Policybazaar Insurance Web Aggregator Pvt Ltd) & Paisabazaar.com  
Approx. 45 TB of Huge Volume Data's AWS Cloud Migration & Transformation:  
Insurance Domain Client Overview:



### PROJECT- 2 INFORMATION ( INSURANCE / SALES DOMAIN )

- **Employment Status:** Permanent
- **Client:** Policybazaar Insurance Web Aggregator Pvt Ltd & Paisabazaar.com

Policybazaar is an Indian insurance aggregator that offers various insurance plans, including life insurance, health insurance, motor insurance, travel insurance, and group plans. The platform facilitates the comparison and purchase of insurance plans based on individual preferences. The company's core technical managemet team has decided to move on-premises data to Amazon Web Services (AWS) cloud.

**MY ROLES & RESPONSIBILITIES :**

As a Lead Data Engineer at Policybazaar, I spearheaded the AWS cloud migration and transformation of approximately 45 TB of insurance data. My key responsibilities included strategic planning, overseeing the data migration process from on-premises to AWS, and ensuring data integrity and security throughout.

- **Data Architect & Engineer:** Designed and built end-to-end data pipelines for efficient data ingestion, transformation, and storage.
- **Data Migration & Pipeline Development:** Led the migration of high-volume vehicle insurance data, transforming it with tools like S3, Kinesis, AWS Lambda, EMR, Glue, and Athena. Implemented real-time streaming data pipelines from MongoDB to DocumentDB, with robust error-handling and backup mechanisms.
- **Data Processing & Transformation:** Designed and optimized Spark applications for data ingestion, and successfully transitioned SQL-based stored procedures into Spark SQL on EMR for improved efficiency and scalability.
- **Data Storage & Database Integration:** Integrated multiple data sources, storing transformed data in S3 in Parquet format for scalability, with an Athena layer for efficient querying. Employed Glue Crawler and DataBrew to automate data cataloging and enhance data quality.
- **Analytics, Reporting & Visualization:** Generated insightful reports using Amazon QuickSight and automated reporting and job scheduling via Apache Airflow. Calculated complex sales revenue and performed in-depth attribute analysis.
- **Performance Optimization & Security:** Ensured optimal performance of big data lakes and cluster configurations while implementing data security best practices across all data processing stages.

This role underscored my ability to manage complex cloud migrations, lead data engineering teams, and deliver scalable solutions within a high-stakes, data-driven environment.

PROJECT - 3 DESCRIPTION



NTTA - Citi Bank – Toll Plaza - (Big Data Engineering)

Insurance & Banking Domain Client Overview:

The North Texas Tollway Authority (NTTA) is an organization that maintains and operates toll roads, bridges, and tunnels in the North Texas area. The NTTA operates toll roads, bridges, and tunnels in the North Texas area and is authorized to acquire, construct, and maintain such projects through toll collection. The customers, both prepaid and postpaid, have Citi Bank accounts.

**MY ROLES & RESPONSIBILITIES:**

- **Data Identification:** Identify the ISSUER database containing billions of attributes and fields such as ACCOUNTNO, CUSTOMER\_DETAILS, CUSTOMERID, VEHICLECLASS, VEHICLENUMBER, REFID, HEXTAGID, GL\_TXNID, GL\_TXNDATE, PAYMENT\_MODE, CURRENT\_BALANCE, ISBLACKLISTED, BLACKLIST\_IN\_DATE, BLACKLIST\_OUT\_DATE, IS\_NEGATIVE\_BALANCE etc.
- **Database Schema Design:** Design the schema for the staging database, identifying the relevant tables for data processing.
- **Data Pipeline Creation:** Create a data pipeline from the client to the staging database for efficient data transfer.
- **Spark Application Development:** Develop a Spark application using Scala for data processing and analysis.
- **Integration with MongoDB:** Integrate Spark with MongoDB to store the processed data in JSON format.
- **Delivery to BI Team:** Deliver the aggregated data in JSON format to the Business Intelligence (BI) team.
- **Report Generation in ELK Stack:** Utilize ELK Stack (Elasticsearch, Logstash, Kibana) for report generation based on the aggregated data.



- **Payroll Company:** Powersoft Global Sol. Pvt. Ltd.
  - **Team Size:** Individual (1) + TPTs
  - **Data Size:** ~ 45 TB
- Technologies and Tools Used:**
- **Data Storage:** AWS S3, MongoDB.
  - **Big Data Processing:** AWS EMR Clusters, Spark-SQL, Hive, ScalaSpark.
  - **Query Execution:** AWS Athena.
  - **Workflow Management:** Apache Airflow.
  - **Reporting:** Power BI, Quicksight.
  - **Programming Language:** Python, Scala.
  - **Database Tools:** Eclipse, MSSQL Server, Dbeaver.
  - **Collaborative Platform:** Linux, Zeppelin.
- In Summary,** this project handling massive data migrations, designing Spark applications, and generating insightful reports for Policybazaar and Paisabazaar. **My individual's role in designing end-to-end pipelines of On-premise to AWS.**

PROJECT- 3 INFORMATION ( INSURANCE & BANKING DOMAIN )

- **Client:** North Texas Tollway Authority (NTTA)
  - **Employment Status:** Permanent
  - **Payroll Company:** Powersoft Global Sol. Pvt. Ltd.
  - **Team Size:** 11
  - **Data Processing:** Kafka, PySpark, Spark Streaming, HDFS, Spark UI, Spark-sql, Hive.
  - **Database Integration:** MongoDB for storing JSON files.
  - **Programming Language:** Scala.
  - **Tools:** Eclipse for development, WinSCP, Putty for server access.
- In Summary,** The Big Data Engineering Project for NTTA involved comprehensive data processing, analysis, and report generation. As a Big Data Analyst, the responsibilities included schema design, data pipeline creation, Spark application development, integration with MongoDB, and delivering aggregated data for report generation using the ELK Stack. The technology stack, including Kafka, Spark, Hive, and MongoDB, reflects a modern and efficient approach to handling large-scale data operations in the toll plaza domain. The project contributes to NTTA's ability to manage customer and vehicle data for improved operational insights and decision-making.

PROJECT - 4 DESCRIPTION

NTTA (Data Science): - Automatic License Plate Recognition (ALPR) or Automatic Number Plate Recognition (ANPR)



Insurance & Banking Domain Client:

Develop a robust ALPR system to accurately identify and track vehicles on North Texas Tollway Authority roads.

**My Contributions:**

- **Data Engineering:** Designed and implemented data pipelines for collecting, cleaning, and preprocessing video data.
- **Computer Vision:** Utilized OpenCV and YOLOv8 for object detection and license plate localization.
- **Machine Learning:** Developed and trained a custom neural network for character recognition.
- **System Integration:** Integrated the components into a real-time pipeline for efficient vehicle tracking and license plate reading.
- **Performance Optimization:** Optimized the system for real-time processing and accuracy.

**Impact:** The developed ALPR system significantly improved toll collection accuracy, traffic management, and security for the North Texas Tollway Authority.



PROJECT- 4 INFORMATION

- **Client:** North Texas Tollway Authority (NTTA)
- **Employment Status:** Permanent
- **Payroll Company:** Powersoft Global Sol. Pvt. Ltd.
- **Team Size:** 11
- **Technologies and Tools Used:**
  - **Image Processing Libraries:** OpenCV for image processing.
  - **Deep Learning Framework:** Keras for building and training neural networks.
  - **Object Detection Algorithm:** YOLOv8 (You Only Look Once, version 8) for license plate detection.
  - **Machine Learning Classifiers:** Multilayer Perceptron (MLP), K Nearest Neighbors (KNN).
  - **EasyOCR:**
  - **Collaborative Notebook:** Jupyter Notebook in Google Colab for collaborative coding and analysis.


PROJECT - 5 DESCRIPTION

MetLife - Intrusion Detection System using Azure Databricks:

Insurance Domain Client Overview:




MetLife is among the largest global providers of insurance, annuities, and employee benefit programs, with 90 million customers in over 60 countries. The Intrusion Detection System (IDS) involves monitoring network activity logs in real-time to enhance the detection of web threats. The primary goal is to generate suspicious activity alerts, support investigations into suspicious activity, and develop network propagation models to identify penetration points.

**MY ROLES & RESPONSIBILITIES:**



PROJECT- 5 INFORMATION ( INSURANCE DOMAIN )

- **Client:** Metlife
- **Payroll Company:** Cognizant Tech. Sol. Pvt. Ltd.
- **Employment Status:** Permanent
- **Technologies:**
  - Azure Databricks: Platform for big data analytics and machine learning.
  - Notebook: Utilized for interactive and collaborative data analysis.

<ul style="list-style-type: none"> <li>• <b>Data Identification &amp; Ingestion:</b> Ingested large Parquet datasets containing network attributes (IP addresses, packets, flags, attack IDs) in Azure Databricks for real-time threat analysis.</li> <li>• <b>Data Enrichment:</b> Enhanced IDS data with additional insights to support in-depth analysis and accurate threat detection.</li> <li>• <b>Exploration &amp; Analysis:</b> Analyzed IDS data to identify attack patterns, trends, and behaviors across network activity.</li> <li>• <b>Model Development:</b> Built and implemented machine learning models to improve intrusion detection accuracy.</li> <li>• <b>Visualization:</b> Created visualizations to display patterns and anomalies in network data, making insights more interpretable for stakeholders.</li> <li>• <b>Results Interpretation:</b> Provided actionable insights into vulnerabilities and threat vectors to support proactive security measures.</li> <li>• <b>Collaboration:</b> Worked closely with security teams to ensure the effectiveness of the IDS.</li> </ul> <p>This comprehensive approach supported MetLife’s global network security by identifying vulnerabilities and enhancing detection capabilities across large-scale data systems.</p>	<p>→ Parquet Data Format: Used for storing and analyzing large datasets efficiently.</p> <ul style="list-style-type: none"> <li>• <b>Programming Language:</b> Scala</li> <li>• <b>Team Size:</b> 16</li> </ul> <p><b>In Summary,</b> the Real-Time Health Data project involves handling a massive scale of health data using technologies such as Kafka, Spark Streaming, and Hive. The responsibilities include collecting, processing, and analyzing health data, with a focus on delivering meaningful insights to the BI team.</p>
PROJECT - 6 DESCRIPTION	PROJECT- 6 INFORMATION ( HEALTH-TECH / HEALTH CARE DOMAIN )
<p><b>Practo - Real-Time Health Data:</b> </p> <p><b>Health Domain Client Overview:</b></p> <p>Practo, a Bengaluru-based health-tech company, has become a comprehensive platform over the past 10 years, offering services such as appointments, consultations, health records, insurance, and online medicine ordering. This project focuses on processing real-time health data using various big data technologies, with a particular emphasis on Spark and Kafka integration.</p> <p><b>MY ROLES &amp; RESPONSIBILITIES:</b></p> <ul style="list-style-type: none"> <li>➤ <b>Data Collection &amp; Integration:</b> Leveraged Spark Structured Streaming and Apache Kafka to ingest and process real-time health data, integrating data from multiple sources, including patient appointments, consultations, health records, and medical orders.</li> <li>➤ <b>Data Processing &amp; Transformation:</b> Employed stateful and stateless transformations, implementing Kafka Sliding Windows and Watermark features to ensure precise data handling and minimize latency. JSON Schema design, handling JSON Using StructType, ArrayType and StructField methods were involved in the coding level with Join operations.</li> <li>➤ <b>Data Storage &amp; Management:</b> Structured data into Avro and CSV formats, optimized for storage and retrieval, and stored in Apache Hive for long-term analysis, ensuring HIPAA compliance and data security.</li> <li>➤ <b>Performance Optimization:</b> Tuned Kafka and Spark configurations to enhance scalability and performance, enabling seamless data flow and reducing processing times.</li> <li>➤ <b>Health Data Analysis:</b> Conducted in-depth analysis on Practo’s data to assess revenue models, patient visits, and doctor orders. Loaded results into Hive for easy access by the BI team.</li> <li>➤ <b>Collaboration &amp; Reporting:</b> Worked closely with the BI team, facilitating data-driven insights through Kibana dashboards and Hive reporting for improved decision-making.</li> </ul> <p>This project showcased expertise in handling high-volume health data, transforming it into actionable insights for Practo, enhancing their operational efficiency and supporting strategic decisions.</p>	<ul style="list-style-type: none"> <li>• <b>Employment Status:</b> Permanent</li> <li>• <b>Client:</b> Practo</li> <li>• <b>Payroll Company:</b> Cognizant Tech. Sol. Pvt. Ltd.</li> <li>• <b>Technologies:</b> Kafka, Spark Structured Streaming, HDFS, Spark UI, Spark-sql, Hive, Kibana</li> <li>• <b>Programming Language:</b> Scala</li> <li>• <b>Team Size:</b> 16</li> <li>• <b>Tools &amp; Platform:</b> Eclipse, SFTP tools.</li> </ul> <p><b>In Summary,</b> the Real-Time Health Data project involves handling a massive scale of health data using technologies such as Kafka, Spark Structured Streaming, and Hive. The responsibilities include collecting, processing, and analyzing health data, with a focus on delivering meaningful insights to the BI team. The comprehensive tech stack, programming in Scala. Hence, such this ways it showcase a sophisticated approach to managing real-time health data for a widely-used health-tech platform.</p>
PROJECT - 7 DESCRIPTION	PROJECT- 7 INFORMATION ( IOT DOMAIN )
<p><b>Opterna – IoT Fiber Optic:</b> </p> <p><b>IoT Domain Client Overview:</b></p> <p>Opterna, an International Fiber Optics Solutions Company, is involved in an Internet of Things (IoT) project where customer sensor devices generate data. The data is then processed and analyzed using a Spark application whether the temperature, humidity etc. are in normal stage or not, with results stored in InfluxDB. Additionally, master and transactional data are stored in PostgreSQL. The overall architecture involves Kafka for data ingestion, Spark for real-time processing, InfluxDB for storage, and PostgreSQL for handling master and transactional data.</p> <p><b>MY ROLES &amp; RESPONSIBILITIES:</b></p> <ul style="list-style-type: none"> <li>➤ <b>Kafka API Development:</b> Developed Kafka APIs to streamline data ingestion from IoT sensor devices, ensuring efficient data transfer into the Kafka ecosystem for analysis.</li> <li>➤ <b>Real-Time Data Processing:</b> Utilized Spark for real-time data processing, analyzing sensor data for temperature, humidity, and other metrics, with results stored in InfluxDB for easy monitoring.</li> <li>➤ <b>Data Storage &amp; Management:</b> Handled master and transactional data storage using PostgreSQL, ensuring robust data organization for IoT insights.</li> <li>➤ <b>Visualization:</b> Designed user-friendly visualization tools to present processed data insights, enhancing operational decision-making.</li> <li>➤ <b>Pipeline Creation:</b> Established end-to-end data pipelines integrating Kafka, Spark, InfluxDB, and PostgreSQL for seamless data flow in IoT operations.</li> </ul> <p>This role highlights expertise in building scalable IoT solutions, leveraging real-time data analysis for impactful insights.</p>	<ul style="list-style-type: none"> <li>• <b>Client:</b> Practo</li> <li>• <b>Payroll Company:</b> Alient Techno Sol. India Pvt. Ltd.</li> <li>• <b>Employment Status:</b> Permanent</li> <li>• <b>Technologies:</b> Kafka, Spark Structured Streaming, HDFS, Spark UI, Spark-sql, Hive, Kibana</li> <li>• <b>Programming Language:</b> Scala</li> <li>• <b>Team Size:</b> 16</li> </ul> <p><b>In Summary,</b> this project involves the storage and processing of reinsurance data, leveraging the capabilities of the Hadoop ecosystem. Your responsibilities encompassed data capture from RDBMS, efficient organization through Hive table management, scripting for data processing, and the optimization of Hive queries for enhanced performance. The integration of Oozie for workflow management ensures the seamless execution of Sqoop and Hive jobs. This comprehensive approach positions HSB to derive meaningful insights from their diverse departmental data for detailed reporting.</p>
PROJECT - 8 DESCRIPTION	PROJECT- 8 INFORMATION ( PHARMA DOMAIN )
<p><b>Pfizer - Pharmaceutical:</b> </p> <p><b>Pharmaceutical Domain Client Overview:</b></p> <p>The Pfizer pharmaceutical project involves the migration of an existing project, which previously utilized MySQL for storing competitor and retailer information, to the Hadoop platform. The primary objectives include the extraction of data from MySQL, storage in an HDFS Data Lake, and the creation of managed and external Hive tables for Business Intelligence (BI) reporting.</p> <p><b>MY ROLES &amp; RESPONSIBILITIES:</b></p> <ul style="list-style-type: none"> <li>➤ <b>Data Migration:</b> <ul style="list-style-type: none"> <li>▪ Developed SQOOP scripts to extract data from MySQL tables and transfer it to the HDFS Data Lake.</li> <li>▪ Ensured the successful re-hosting of data from MySQL to Hadoop.</li> </ul> </li> <li>➤ <b>Hive Table Management:</b> <ul style="list-style-type: none"> <li>▪ Created managed and external Hive tables on top of the extracted data.</li> <li>▪ Facilitated BI team report generation through organized data structures.</li> </ul> </li> <li>➤ <b>Hive Query Development and Optimization:</b> <ul style="list-style-type: none"> <li>▪ Developed data queries using Hive Query Language (HQL).</li> <li>▪ Optimized Hive queries for improved performance, handling complex SQL queries efficiently.</li> </ul> </li> <li>➤ <b>Data Processing and Ad-hoc Requests:</b></li> </ul>	<ul style="list-style-type: none"> <li>• <b>Employment Status:</b> Permanent</li> <li>• <b>Client:</b> Pfizer</li> <li>• <b>Payroll Company:</b> Alient Techno Sol. India Pvt. Ltd.</li> <li>• <b>Technologies:</b> Hadoop, Apache Hive, Sqoop, Apache Pig, Cloudera</li> <li>• <b>Team Size:</b> 7</li> </ul>

- Performed data ingestion tasks using Hive, ensuring data integrity and availability.
- Addressed ad-hoc requests by developing and executing tailored solutions using Hive.

## PROJECT - 9 DESCRIPTION

## PROJECT-9 INFORMATION ( BANKING DOMAIN )

### U.S. BANK - Banking:



#### Banking Domain Client Overview:

U.S. Bank is a leading financial institution offering a wide range of banking and financial services. It serves individuals, businesses, and institutions across the United States, providing products and services such as checking and savings accounts, loans, credit cards, investment services, and wealth management.

#### MY ROLES & RESPONSIBILITIES:

- Data Import & Export:** Managed data transfer to and from HDFS, Hive, and relational databases using Sqoop, enabling seamless data flow.
- Performance Tuning:** Optimized performance by enhancing scalability, adjusting batch intervals, and tuning Spark applications for efficient processing.
- Hive Table Management:** Created and loaded Hive tables for structured data storage and enabled efficient query operations.
- Spark Development:** Developed Scala scripts for data aggregation using DataFrames and RDDs, handling large data sets with advanced Spark capabilities.
- Algorithm Optimization:** Improved Hadoop algorithms using Spark Core and Spark SQL, integrating advanced text analytics.
- Technical Specifications:** Drafted technical designs for ETL processes, ensuring thorough documentation and accuracy.
- Code Review:** Conducted code reviews and bug fixing for quality and performance improvement.

This project emphasized leveraging big data technologies to optimize KPIs and drive insights for U.S. Bank's progress across demographic areas.

- Client:** U.S. BANK
- Payroll Company:** Alient Techno Sol. India Pvt. Ltd.
- Employment Status:** Permanent
- Team Members:** 5
- TECHNOLOGIES:** MapReduce, Hive, Sqoop, SparkSQL HDFS, Hue
- Programming Language:** Java & Scala
- TOOLS & Platform:** IntelliJ, WinSCP, Putty, CDH 5.7.0
- Cluster Information:** 90 Nodes cluster of 1.5TB Capacity of each node

## PROJECT - 10 DESCRIPTION

## PROJECT-10 INFORMATION ( TELECOM DOMAIN )

### (A) TELECOMMUNICATION CDR DATA ANALYSE AND PROCESS USING APACHE HADOOP TECHNOLOGIES:



#### Telecom Operator Domain Client Overview:

- Digicel Jamaica is a premium telecom operator covering 37 regions in Central America.
- It offers a range of niche products/services, making it the largest operator in the region.
- The subscriber base is extensive, indicating a strong market presence.

#### MY ROLES & RESPONSIBILITIES:

- CDR System Overview:**
  - CDRs are generated in a telecom network when one party calls another.
  - Information captured in CDRs is crucial for tracking Key Performance Indicators (KPIs) such as duration, chargeable amount, roaming data, and more.
  - CDRs are generated in various file formats, including CSV, XML, and JSON.
  - Data is organized on a day-by-day basis.
- CDR Properties:**
  - Properties include information on whether the call is incoming or outgoing, whether it's in roaming, chargeable amount, duration, and network details.
  - Other dimensions like day key, event category code, call duration code are used for analysis.
- Storage and Database:**
  - CDR information is stored in different file formats.
  - Code-related information is stored in a Relational Database Management System (RDBMS) as individual tables.
  - Look-up tables in the RDBMS are used to map with the base CDR information.
- Key Data Points in CDRs:**
  - Incoming/outgoing status of calls.
  - Roaming status of the call.
  - Chargeable amount for the call.
  - Duration of the call.
  - Network details, including whether the call is within the same network or in another.
- Usage of Look-Up Tables:**
  - Look-up tables in the RDBMS are utilized to store and retrieve code-related information.
  - These tables help in mapping and associating information from CDRs with the relevant code details.

Involved in designing Complex Hive Queries(HQL) to Analyse Data as per the Requirement KPI & Dimensions. Process Analysed Data into Resultant Hive Internal Tables Using Hive HQL, MapReduce, Oozie Workflow Schedule in Hue Environment. Hive-HBase Integration.

### (B) HIVE COLUMN LEVEL SECURITY:

#### Secure and Load Sensitive Data by Encryption and Retrieve Data After Decryption:

#### RESPONSIBILITIES:

- Encrypting sensitive data and loading it into Hadoop Distributed File System (HDFS) using Apache Hive User-Defined Functions (UDF), and subsequently decrypting the data using the AES algorithm from the Java Cryptography Architecture (JCA).

- Client:** Digicel Jamaica
- Payroll Company:** Hinx Technologies Pvt. Ltd.
- Employment Status:** Permanent
- Team Members:** 7
- Technologies:** MapReduce, Hive, Sqoop, Oozie, Zookeeper, HDFS, HBase, Kafka, Flume, Hue
- Database:** HBase, Oracle
- Programming Language:** Java
- TOOLS & Platform:** Eclipse IDE, WinSCP, Putty, CDH 5.7.0

In summary, Digicel Jamaica's telecom infrastructure relies on a robust CDR system to capture and analyze crucial call-related information. The use of different file formats and an RDBMS with look-up tables enhances the efficiency of managing and extracting valuable insights from the generated data.

