

Bigdata Architecture for Realtime Fraud Detection

- By Kausik Chattapadhyay
- 02/14/2024
- DSC650-Big Data
- Prof. Nasheb Ismaily



Agenda



PROBLEM
STATEMENT



PROPOSED
ARCHITECTURE



DATA INGESTION



DATA
TRANSPORTATION
AND BUFFERING



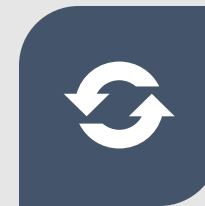
DATA STORAGE AND
MANAGEMENT



DATA PROCESSING
AND ANALYSIS




MACHINE LEARNING
MODEL DEPLOYMENT



FEEDBACK LOOP FOR
MODEL
IMPROVEMENT

Problem Statement

Challenge: The bank faces escalating threats of credit card fraud and increasing regulatory pressure for anti-money laundering (AML) measures.



Impact: Current systems are inadequate, leading to financial losses, regulatory fines, and reputational damage.

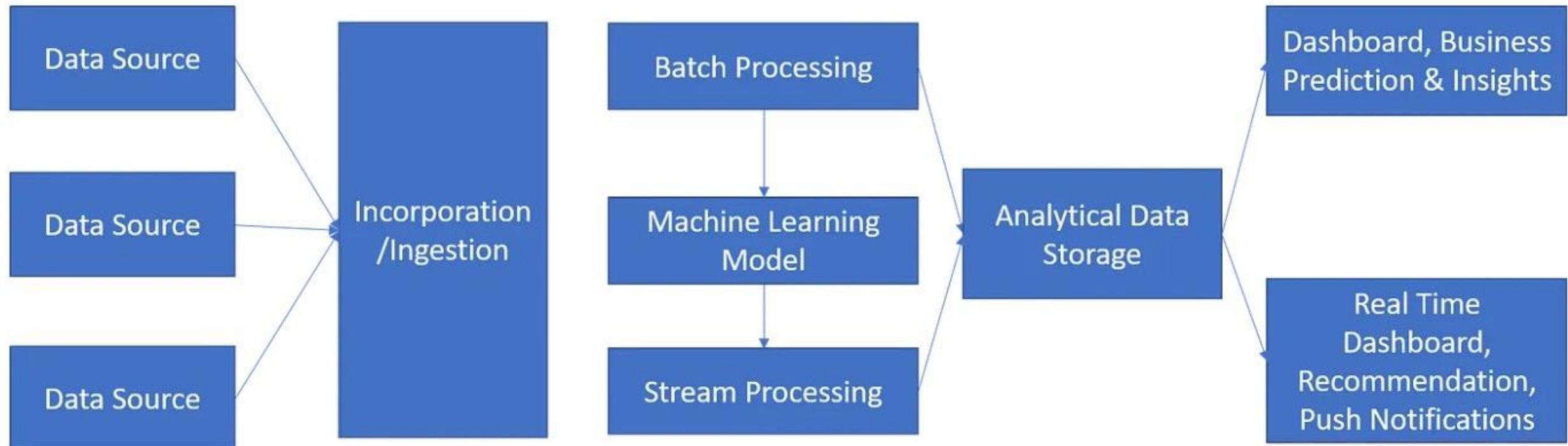


Objective: Develop a robust solution to detect and prevent fraud in real-time while ensuring compliance with AML regulations.



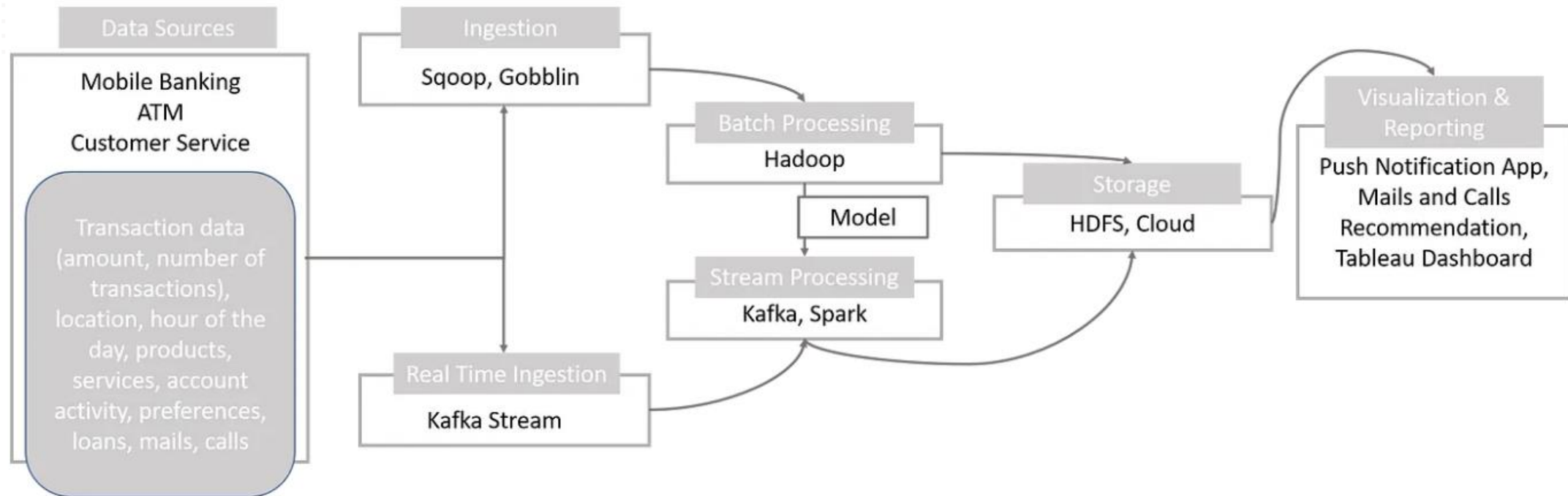
Focus: Utilize historical and real-time data to enhance fraud detection capabilities and minimize risks.

Proposed Architecture



Data Flow and Analysis

- Transactional data from websites, mobile apps, and ATM machines flows into Kafka for real-time streaming.
- Spark conducts immediate fraud detection by processing streaming data, utilizing feature selection and machine learning models.
- Quick response actions, such as notifications and alerts, are triggered to mitigate fraudulent activities swiftly.



Key Points



Data ingestion from various sources, including website, mobile banking application, ATM machine, and bank's database, was achieved using Sqoop.



Real-time processing utilized Gobblin, while batch processing was handled by Kafka Stream.



Data was then transferred to Kafka for real-time analytics or HDFS of Hadoop before being processed in Spark.



Different data storage solutions : data warehouses for structured data, data lakes for diverse data types, NoSQL databases for unstructured data and high throughput applications, and in-memory databases for low-latency access.

Key Points



Anomaly detection in fraud prevention include One-Class SVM, DBSCAN, and Isolation Forest, while the batch-trained model will be used to analyze data in the stream part for building real-time dashboards, recommendations, and notifications.



Generating results for building interactive real-time dashboards, recommendations, and notifications.



Automated feedback loops in fraud detection involved reducing false positives, leveraging negative data, and updating systems with insights from risk analysts to enhance detection and prevention.



Amazon SageMaker can effectively implement feedback loops in machine learning systems for bank fraud detection, continuously improving model accuracy, adapting to evolving fraud patterns, and ensuring compliance with regulations.

Benefits and Conclusion

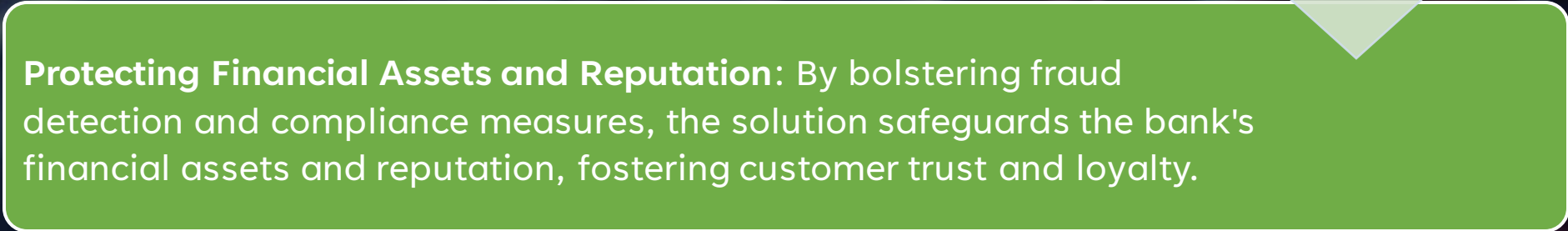
Enhanced Fraud Detection: Real-time analytics powered by Spark enable swift identification of fraudulent activities, minimizing financial losses.




Ensured Compliance: Comprehensive data storage and analysis capabilities ensure adherence to AML regulations, mitigating the risk of regulatory fines.



Protecting Financial Assets and Reputation: By bolstering fraud detection and compliance measures, the solution safeguards the bank's financial assets and reputation, fostering customer trust and loyalty.





”

Thank you