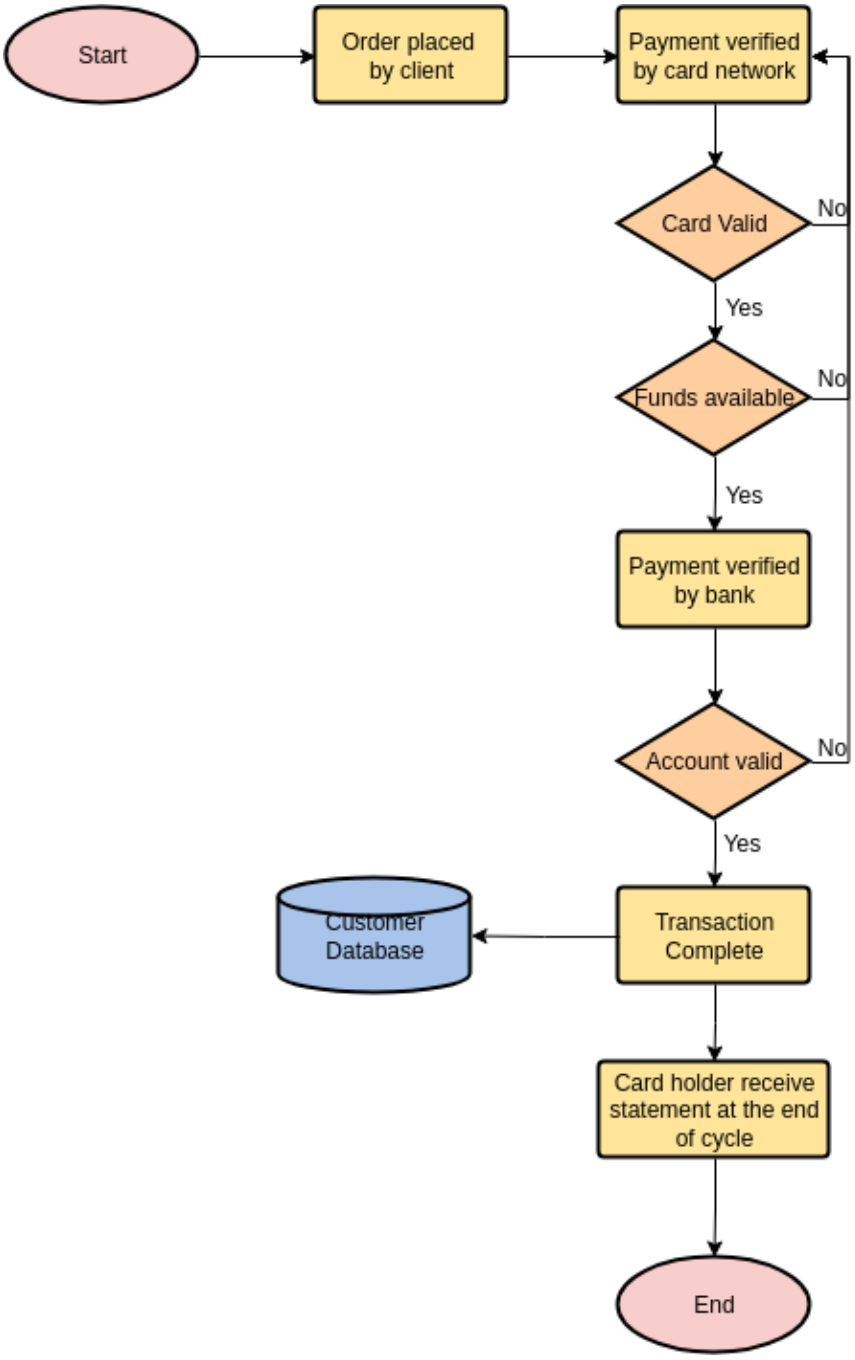
Case study task 1: Team Sigma

Client:Natwest credit card fraud detection case study:

Credit card fraud is a significant concern for financial institutions and cardholders. Detecting fraudulent transactions in real-time is crucial to prevent financial losses. The Hadoop ecosystem provides a powerful platform for processing large volumes of data and implementing advanced analytics algorithms, making it an ideal choice for credit card fraud detection.

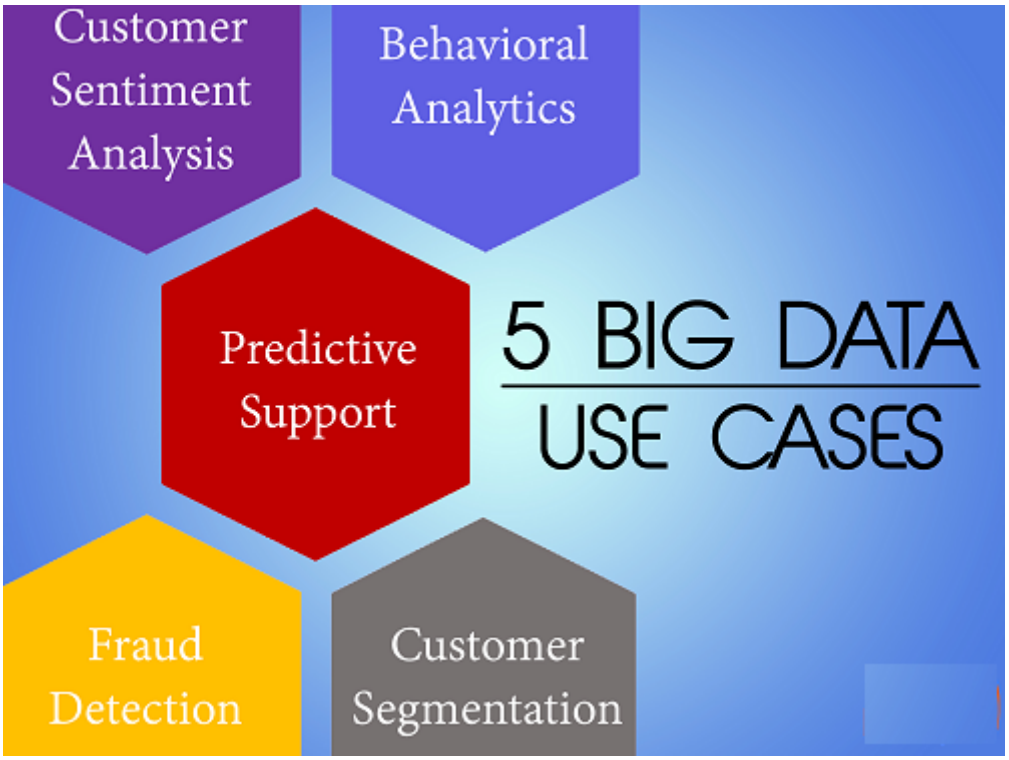
How this system works:



Problem case:

The objective of this case study is to develop a credit card fraud detection system using the Hadoop ecosystem. The system should be able to process large volumes of transaction data in real-time and identify potentially fraudulent transactions accurately.

Solution Overview: Hadoop ecosystem components used to achieve the result.



The solution will leverage various components of the Hadoop ecosystem, including HDFS (Hadoop Distributed File System), MapReduce, Hive, and Spark, to build an end-to-end credit card fraud detection system.

1. **Behavioral Analytics**- An abnormal number of clicks from the same IP address or a pattern in the access times — although this is the most obvious and easily identified form of click fraud, it is amazing how many fraudsters still use this method, particularly for quick attacks. They may choose to strike over a long weekend when they figure you may not be watching your log files carefully, clicking on your ad

repeatedly so that when you return to work on Tuesday, your account is significantly depleted. Part of this fraud might be unintentional when a user tries to reload a page.

2. **Data Ingestion:** Transaction data from various sources, such as credit card processors or banking systems, will be ingested into the Hadoop cluster. This data will be stored in HDFS (data servers for handling large volumes of data) for further processing.

3. **Data Preprocessing:** The raw transaction data will undergo preprocessing steps to clean and transform it into a suitable format for analysis. This may include removing duplicates, handling missing values, and normalising numerical features so that we get a clean data output for further processing.

4. **Feature Engineering:** Relevant features will be extracted from the transaction data to capture patterns indicative of fraudulent activity. Features like transaction amount, location, time of day, merchant category code (MCC), etc., can be used. The system will also read the behaviour of the uses of the card at different locations and frequency.

5. **Training our system to read data:** Using functions to create a fraud detection model using historical transaction data labelled as either fraudulent or legitimate. Techniques like pattern reading and decision trees can be employed.

6. **Real-time Processing:** As new transactions arrive in real-time, they will be processed by applying the trained model to predict the likelihood of fraud. This can be done using Spark Streaming or Apache Flink for real-time data processing.

7. **Predictive support using Alert Generation:** Transactions identified as potentially fraudulent will trigger alerts or notifications to relevant stakeholders, such as cardholders, financial institutions, or fraud detection teams.

8. **Reporting and Visualization:** The system will provide comprehensive reporting and visualisation capabilities to monitor fraud detection performance, generate insights, and identify emerging patterns using a dashboard or any other easy to understand tools.

Benefits

Implementing a credit card fraud detection system using the Hadoop ecosystem offers several benefits:

1. **Scalability:** Hadoop's distributed computing framework allows processing large volumes of transaction data efficiently, enabling real-time fraud detection even with high transaction volumes.

2. **Flexibility:** The Hadoop ecosystem provides a wide range of tools and libraries that can be leveraged for data preprocessing, feature engineering, model training, and real-time processing. This flexibility enables customization based on specific business requirements.

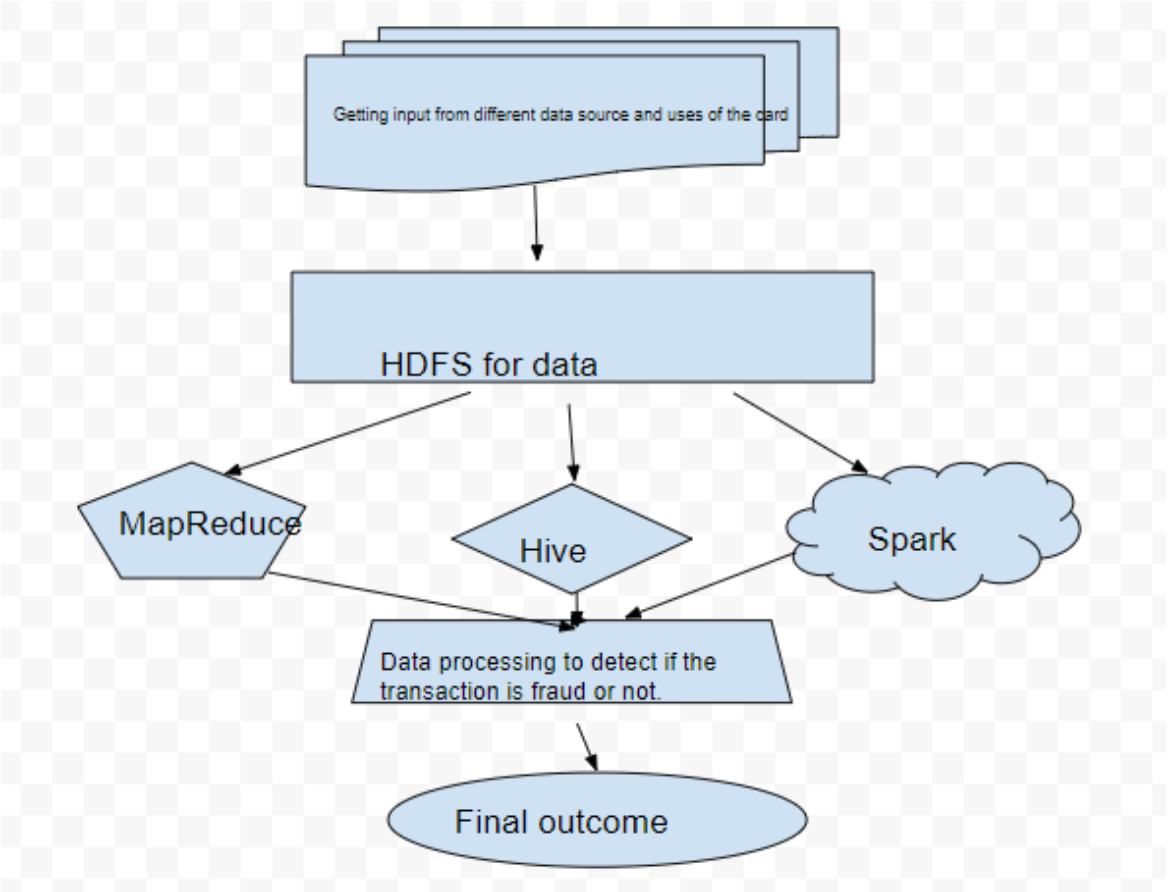
3. **Cost-effectiveness:** Hadoop's open-source nature eliminates the need for expensive proprietary software licences. Additionally, its ability to run on commodity hardware reduces infrastructure costs compared to traditional solutions.

4. **Real-time Detection:** By leveraging streaming technologies like Spark Streaming or Apache Flink, the system can detect fraudulent transactions in near real-time, minimising potential losses.

Conclusion

Credit card fraud is a significant challenge faced by financial institutions worldwide. Leveraging the power of the Hadoop ecosystem enables the development of an efficient and scalable credit card fraud detection system. By combining data preprocessing, feature engineering, machine learning algorithms, and real-time processing capabilities, this solution can help identify fraudulent transactions promptly and mitigate financial losses effectively.

Working flow chat for the system:



Showing graphical representation: