Case Study: Real-time advanced analytics (ML/Big Data)

Abstract and learning objectives

Woodgrove Bank, who provides payment processing services for commerce, is looking to design and implement a proof-of-concept (PoC) of an innovative fraud detection solution. They want to provide new services to their merchant customers, helping them save costs by applying machine learning and advanced analytics to detect fraudulent transactions. Their customers are around the world, and the right solutions for them would minimize any latencies experienced using their service by distributing as much of the solution as possible, as closely as possible, to the regions in which their customers use the service.

In this whiteboard design session, you will work in a group to design the data pipeline PoC that could support the needs of Woodgrove Bank.

At the end of this workshop, you will be better able to implement solutions that leverage the strengths of Cosmos DB in support of advanced analytics solutions that require high throughput ingest, low latency serving and global scale in combination with scalable machine learning, big data and real-time processing capabilities.

Step 1: Review the customer case study

Outcome

Analyze your customer's needs.

Timeframe: 15 minutes

Directions: With all participants in the session, the facilitator/SME presents an overview of the customer case study along with technical tips.

1. Meet your team members and trainer.

2. Read all directions for steps 1-3 in the student guide.

3. As a team, review the following customer case study.

Customer situation

Woodgrove Bank, who provides payment processing services for commerce, is looking to design and implement a PoC of an innovative fraud detection solution. They know from experience and through contacts in the financial industry that there is a constant arms race between fraudsters and banks. Thanks to increasingly powerful and easily accessible technology, financial crime is on the rise. Payment processing companies, like Woodgrove Bank, and their merchant customers risk financial losses due to fraud.

They also risk fines from failing to detect or even prevent criminal acts like money laundering or terror financing. Woodgrove forecasts reaching over USD $10 Billion in assets over the upcoming fiscal year, placing them within the stricter regulatory purview of institutions classified by the US government as "big banks". This means that they will be subject to regulatory fines over and above the fraud loss, putting their business at greater risk.

While all forms of fraud are on the rise, like ATM fraud, card transaction fraud, payment fraud, Woodgrove Bank would like to focus on online fraud. In the most basic terms, online fraud is committed when an unauthorized user impersonates another user by taking over their account, using malware, or hijacking internet sessions and uses the impersonated credentials to make purchase transactions. When dealing with millions of transactions, it is both crucial and challenging to detect and monitor fraud in real-time across all transactions. Doing so helps prevent additional losses and detect widespread attacks.

Given this focus on online fraud, they want to provide new services to their merchant customers, helping them save costs by applying machine learning and advanced analytics to detect fraudulent transactions. Their customers are around the world, and the right solutions for them would minimize any latencies experienced using their service by distributing as much of the solution as possible, as closely as possible, to the regions in which their customers use the service. This is the solution for which they would like to implement a PoC.

In flagging fraudulent transactions, they know there are tradeoffs between being overly aggressive and mistakenly identifying innocuous transactions as fraudulent, and not being aggressive enough such that they miss transactions that represent real fraud. According to Mari Stephens, Chief Information Officer (CIO), Woodgrove Bank, they would rather miss a fraudulent event in their automated system, than mistakenly identify innocuous transactions as fraudulent because the latter will frustrate both their merchant customer and the end customers and potentially lose their business. However, they want to balance this by doing as much as they can to detect fraud while minimizing the customer frustration. To address this, they believe the PoC will need to handle transactions at two "speeds". First, they want to screen transactions for fraud as they happen, only blocking a transaction if the system is very confident it is fraudulent. Second, they want to perform a more in-depth, offline fraud sweep of transactions to identify suspicious transactions. These are transactions which are potentially fraud, for which they will notify the merchant that they should perform additional verification with the end customer before completing the order. This deeper analysis that is performed in batch may use a slightly different ML model, but since it is a more intensive run, it needs to be run in batch and score transactions a little less leniently than the real-time scoring model. Remember, Woodgrove wants to minimize false positives during real-time scoring, but do a deeper analysis of the transactions later on and possibly tag those that were not blocked as suspicious.

They have decades worth of historical transaction data (including transactions identified as fraudulent) that they believe would be helpful in the fraud detection PoC. This data is in tabular format and can be exported to CSV files if needed.

The analysts at Woodgrove Bank are very interested in the recent notebook-driven approach to performing data science and data engineering tasks, and would prefer a solution that features notebooks as the standard way to explore data, prepare data, model, and define the logic for scheduled processing.

Woodgrove's current process

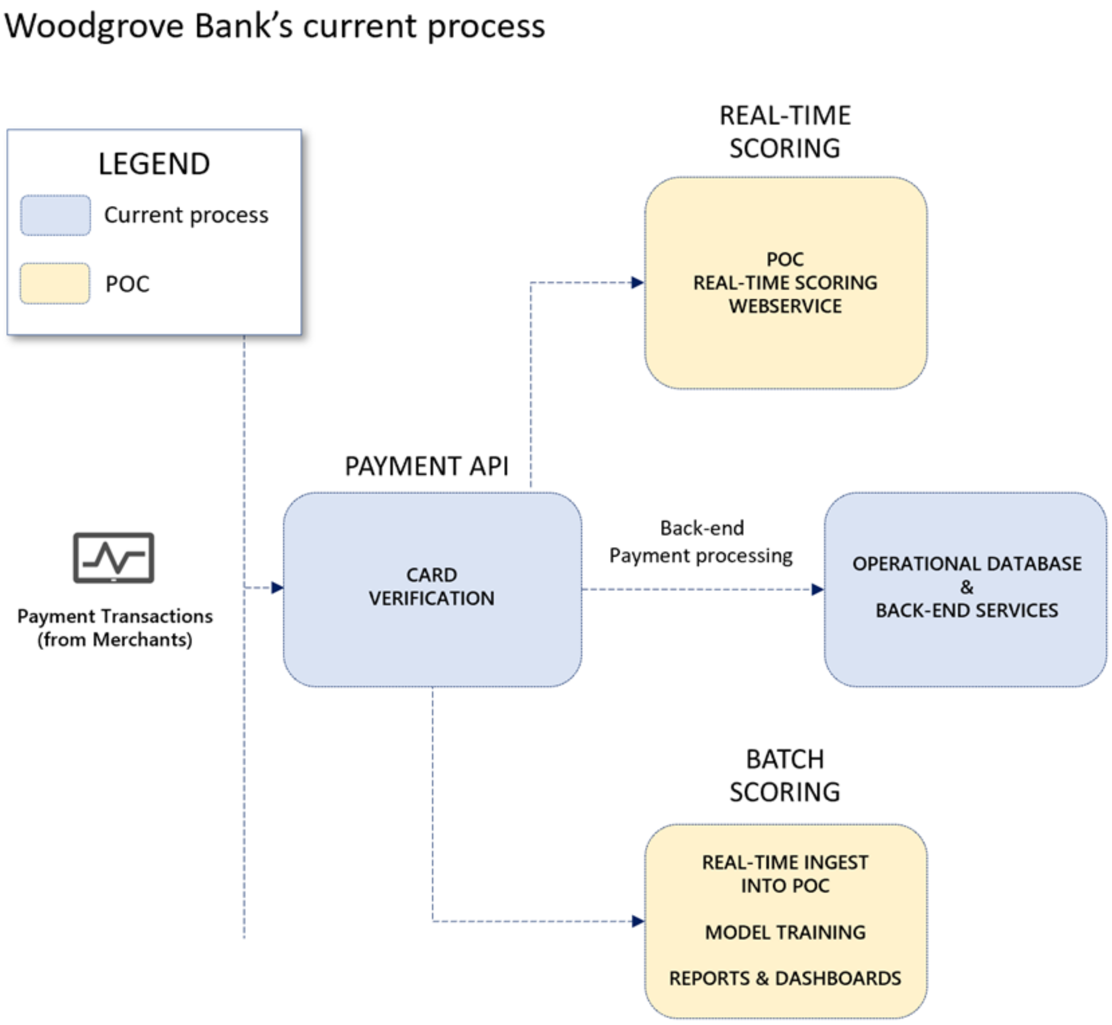
Woodgrove Bank provides a RESTful API that their merchant customers use to submit payments. The POC you design should not interrupt this process in any way. The solution you design needs to run side-by-side and augment their current process without changing their current workflow. Currently, as payments flow through their API endpoints, a series of cardholder verification steps are executed, such as matching the cardholder's billing address to their account. Once this validation check has completed, Woodgrove returns an authorization ID to the merchant, along with a status (accepted, rejected, declined, etc.). The payment details are entered into a relational database and the back-end payment process continues. There may be an opportunity to modify this process down the road, but that is not the focus of the POC.

The customer is asking for 2 additions to their current process:

- A RESTful API that can be called for immediate scoring on a transaction to see whether it should be blocked due to reasonably high-level confidence that it is fraudulent. Remember, this step should have a low number of false positives. The batch process that conducts a deeper sweep should flag suspicious transactions that were not blocked by this initial check.
- A real-time data ingestion pipeline they can pass data to at the time they save the payment transaction data from within their API. This should sit side-by-side with their current process, not change it.

To clarify, the requirement for real-time scoring of the payment transaction as fraudulent is not the same as the real-time ingest of all payment transaction data.

Below is a simple diagram Woodgrove Bank provided of their current process (blue boxes), showing where they would like you to fit in the new POC components (yellow boxes).


Woodgrove Bank's current process
LEGEND
Current process
POC
REAL-TIME SCORING
POC REAL-TIME SCORING WEBSERVICE
PAYMENT API
CARD VERIFICATION
Back-end Payment processing
OPERATIONAL DATABASE & BACK-END SERVICES
Payment Transactions (from Merchants)
BATCH SCORING
REAL-TIME INGEST INTO POC
MODEL TRAINING
REPORTS & DASHBOARDS

Customer needs

1. Need to provide fraud detection services to our merchant customers, using incoming payment transaction data to provide early warning of fraudulent activity.

2. We would like to schedule offline scoring of "suspicious activity" using our trained model to create aggregates showing statistics around detected fraudulent activity, and make that data globally available in regions closest to our customers through our web applications.

3. For all transactions flowing through our system, we want to use our trained model to make near real-time predictions of fraudulent activity.

4. We want the ability to analyze all transactions over time, so we need to be able to store data from transaction sources into long-term storage, without interfering with jobs reading the data set.

5. We would like to use a standard platform that supports our near-term data pipeline needs while providing a long-term standard for data science, data engineering, and development.

Customer objections

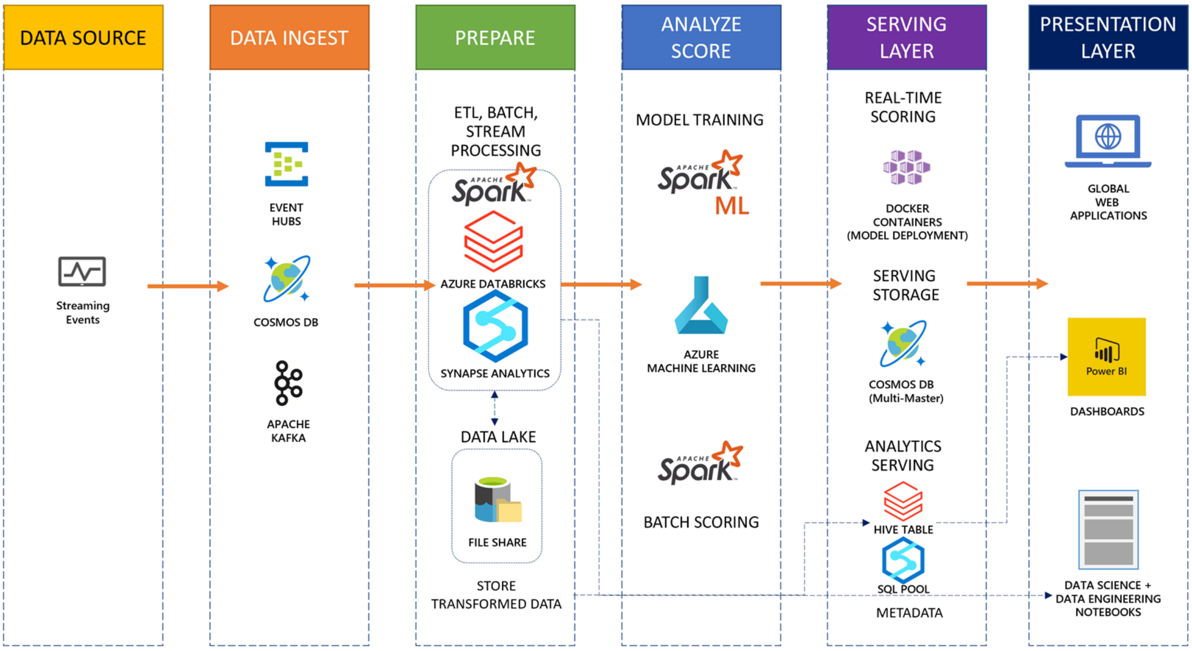
1. It's not clear to us if we can only use Cosmos DB as our web app's database, or if we should consider using it in other parts of our advanced analytics data pipeline such as for real-time transaction ingest or for serving of offline processed data.

2. Does Cosmos DB integrate with open source big data analytics like Apache Spark?

3. Properly selecting the right algorithm and training a model using the optimal set of parameters can take a lot of time. Is there a way to speed up this process?

4. We are concerned about how much it costs to use Cosmos DB for our solution. What is the real value of the service, and how do we set up Cosmos DB in an optimal way?

Infographic for common scenarios



Step 2: Design a proof of concept solution

Outcome

Design a solution and prepare to present the solution to the target customer audience in a 15-minute chalk-talk format.

Timeframe: 60 minutes

Business needs

Directions: With your team, answer the following questions and be prepared to present your solution to others:

1. Who will you present this solution to? Who is your target customer audience? Who are the decision makers?
2. What customer business needs do you need to address with your solution?

Design

Directions: With your team, respond to the following questions:

High-level architecture

1. Without getting into the details (the following sections will address the particular details), diagram your initial vision for handling the top-level requirements for payment fraud detection, including stream capture and

processing, long-term storage, model training, global distribution of the model for real-time scoring and of the pre-scored fraud data, and dashboards.

Globally distributed data

1. Which data storage service would you recommend for storing the suspicious transactions? Remember, Woodgrove Bank wants to minimize access latency for their global customers. Be specific about how data is replicated.

2. How does your chosen service handle scaling to meet varying levels of demand across different regions? Can you set specific capacity for specific regions?

3. Distributed databases that replicate data to multiple locations have some potential delay between when you write a record and when that record is available for reading. What options does your chosen service have to ensure the data is not "stale" when read? Are there any tradeoffs between reducing the window between writes, and if so, how do they apply to Woodgrove Bank's situation?

Data ingest

1. What are your recommended options for ingesting payment transaction events as they occur in a scalable way that can be easily processed while maintaining event order with no data loss?

2. Of the ingest options you identified previously, which would you recommend for the scenario?

Data pipeline processing

1. Woodgrove Bank indicated that they would like a unified way to process both streaming data and batch data on a platform that can also support their data science, data engineering, and development needs. Which platform would you recommend, and why?

2. The big data systems Woodgrove Bank used in the past were only able to append new data to the end of existing data sets. This meant each time they had to update, they would actually create a duplicate row containing the changed data and then have to author queries to merge those rows so that they had a clean view of the current state of the data. How will your chosen platform cope with this challenge?

3. How will your chosen data processing platform connect to and process data from your chosen data ingest solution for streaming data?

4. What configuration would you need to apply to your solution to allow it to restart any stream processing in the case the job is stopped?

5. What specific secrets might their processing solution want to store? How would they securely store and access those secrets?

Long-term data storage

1. As incoming data is processed, refined, and scored, all of the transactions need to be persisted to long-term storage for analysis, model training and validation, and reporting. This storage needs to handle long-term growth, be fast enough to rapidly ingest new data while simultaneously handling reads against the same data set without interference, and act as a reliable data source for dashboards and reports. Which is your recommended long-term data storage solution, keeping in mind its role within your selected data pipeline processing platform?

Model training and deployment

1. Describe how your chosen data processing platform will support machine learning model training and deployment. The model will need to be trained on and validated against historical payment transaction data that includes known fraudulent transactions.

2. How will you schedule regular batch scoring of fraud data using the trained model, and make that data available to Woodgrove Bank's web applications at a global scale?

Dashboards and reporting

1. Woodgrove Bank's business analysts would like to have a set of dashboards they can monitor that provide real-time views of fraud trends at a global scale. Thinking back to how your proposed solution provides a set of summary tables containing business-level aggregates, what do you propose using to meet this requirement? Be specific about how this solution will be put in place and which features it supports.

2. Woodgrove Bank's data analysts, who build and maintain reports, are comfortable working with T-SQL. How can they efficiently access the data for analytical queries, ensuring they have access to the most up-to-date data, without impacting the transactional data store?

Prepare

Directions: As a team:

1. Identify any customer needs that are not addressed with the proposed solution.
2. Identify the benefits of your solution.
3. Determine how you will respond to the customer's objections.

Prepare a 15-minute chalk-talk style presentation to the customer.

Step 3: Present the solution

Outcome

Present a solution to the target customer audience in a 15-minute chalk-talk format.

Timeframe: 30 minutes

Presentation

Directions:

1. Pair with another team.
2. One group is the Microsoft team, the other is the customer.
3. The Microsoft team presents their proposed solution to the customer.
4. The customer makes one of the objections from the list of objections.
5. The Microsoft team responds to the objection.
6. The customer team gives feedback to the Microsoft team.
7. Switch roles and repeat Steps 2-6.