Real world whiteboard design case study

Vending Machines

Microsoft

Step 1:
Review the customer case study

Outcome
Analyze your customer's needs

Timeframe
15 minutes

Customer situation
Trey Research

Trey Research Inc. looks at the old way of doing things in retail and introduces innovative experiences that delight customers and drive sales.

Their latest initiative focuses on intelligent vending machines that have sophisticated computing platforms called Vending Machine IO boards (which are capable of running Microsoft Windows)

These boards are well advanced beyond the traditional Vending Machine Controller (VMC) that are only capable of rudimentary functions like controlling temperature, dispensing product and processing cash payment.

Each vending machine includes a large, hi definition display, a touch screen, a camera, and peripherals for handling cash and credit cards, and they are all connected via either WiFi or 4G LTE connections to the Internet.

Customer situation

- Looking at designing a solution that addresses three core areas: commerce, engagement analytics, and intelligent promotions.
- Commerce: they are looking at modernizing the handling of purchase transactions that accept payment by cash or credit card, and in the future, NFC (Near Field Communication). Regardless of the payment method or the transaction outcome, all purchase transactions are logged to a database for later analysis.
- Engagement: Collect telemetry like dwell times, impressions, and conversions to answer questions like when a recommendation is displayed and a purchase transaction occurred, was the recommended product purchased or not

Customer situation

- Intelligent Promotion: Each vending machine maintains a local copy of the ads (images, videos) that are appropriate to the inventory and used with promotions. These are pushed as a package down to the vending machine on demand. When a vending machine identifies a visitor in proximity, it takes a photo that it will use to anonymously determine demographics (such as age, gender, and possibly features like whether the consumer is smiling or wearing sunglasses) that are used to decide what to promote on its display.

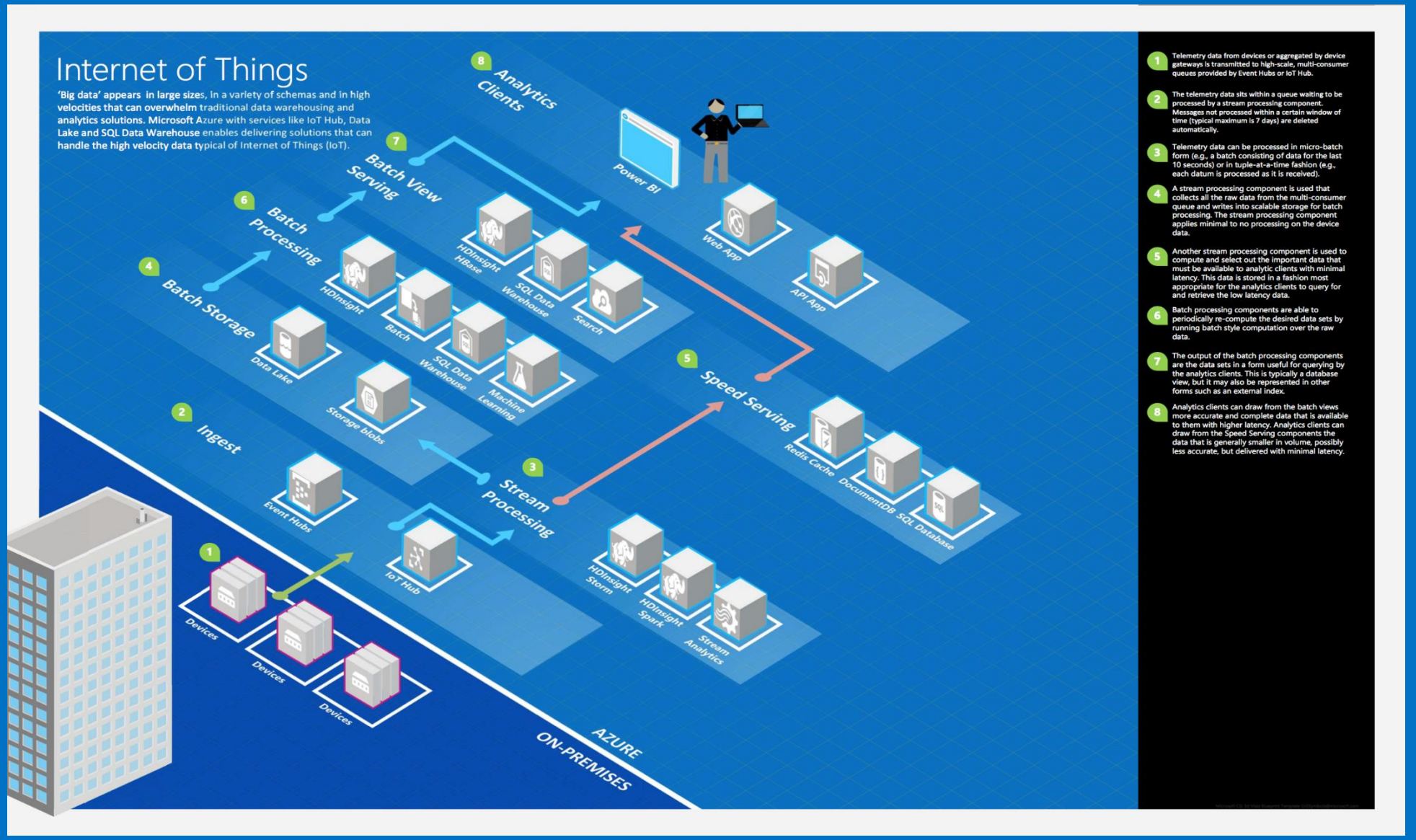
Customer needs

- An IoT solution that can handle high volumes of telemetry data, and enables the solution to communicate with the vending machines for situations like package updates.
- A data store that can handle the extremely write-heavy workload that results from the purchase transactions, whilst still allowing them to quickly perform analytics using SQL.
- A platform on which to build and train machine learning models against high volumes of training data, ideally programmed with R.
- A solution that can provide demographics, given a photo of a person.
- A highly scalable storage solution that won't "max out" and can handle all the telemetry from their vending machines.
- Tools for performing light-weight wrangling of their data, exploration and visualization.

Customer objections

- We've heard that Azure's machine learning can only train on data sets up to 10GB in size, are we blocked?
- While not required in the short term, would our machine learning approach enable us to support reinforcement learning (whereby recommendations that lead to a purchase are preferred over time)?
- Can we really perform real-time analytics using only a single data store? We cannot afford to lose any of our purchase transaction data.
- We are concerned that training our models will take too long.

Common scenarios



Step 2:
Call to action: Design the solution

Outcome

Design a solution and prepare to present the solution to the target customer audience in a 30-minute chalk-talk format.

Timeframe

60 minutes

Business needs (10 minutes)	• Respond to questions outlined in your guide and list the answers on a flipchart.
Design (35 minutes)	• Design a solution for as many of the stated requirements as time allows. Show the solution on a flipchart.
Prepare (15 minutes)	• Identify any customer needs that are not addressed with the proposed solution. • Identify the benefits of your solution. • Determine how you will respond to the customer's objections. • Prepare for a 10-minute presentation to the customer.

Step 3:
Call to action: Present the solution

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

Timeframe
30 minutes (15 minutes one team, to present and receive feedback.)

Directions
- Pair with another table.
- One table is the Microsoft team and the other table is the customer.
- The Microsoft team presents their proposed solution to the customer.
- The customer asks one of the objections from the list of objections in the case study.
- The Microsoft team responds to the objection.
- The customer team gives feedback to the Microsoft team.

Microsoft

Wrap -up

Outcomes

- Identify the preferred solution for the case study.
- Identify solutions designed by other teams.

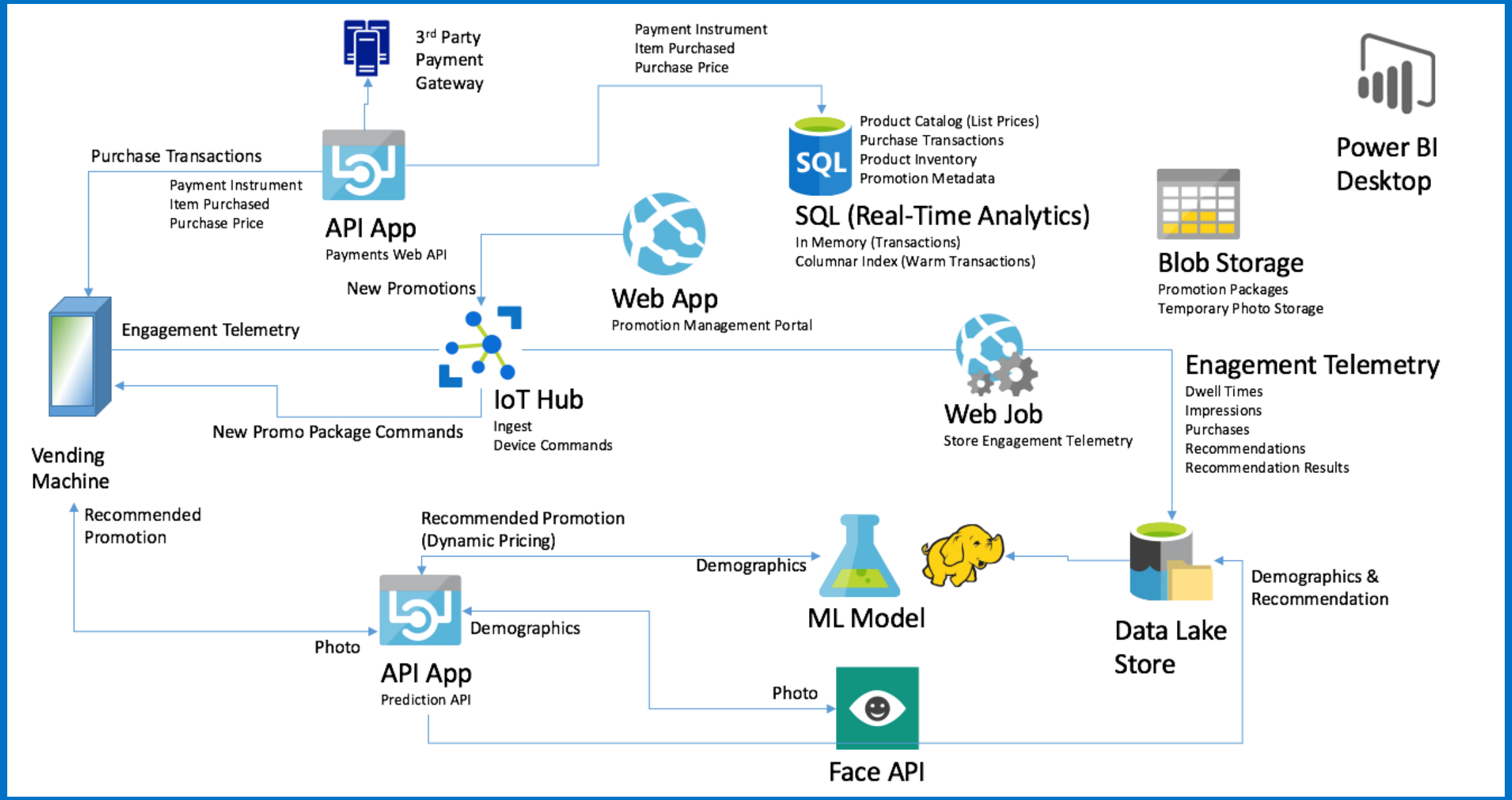
Timeframe

15 minutes

Preferred target audience

- Grant Trey, Chief Innovation Officer, Trey Research
- The primary audience is business decision makers and technology decision makers.
- Usually we talk to the Infrastructure Managers who report into the CIOs, or to application sponsors (like a VP LOB, CMO) or to those that represent the Business Unit IT or developers that report into application sponsors.

Preferred solution



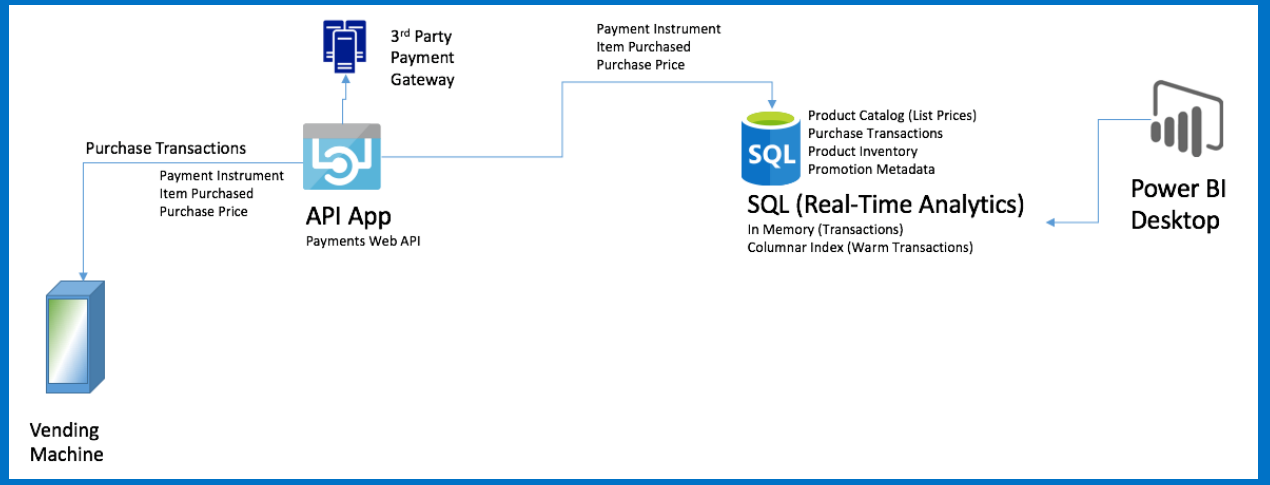
Preferred solution

Commerce

How would you recommend that Trey complete purchase transactions and store their purchase transaction history in Azure?

Trey research should provision an API App that will host the web service responsible for payments, make a call out to the 3rd party payment gateway to authorize and capture payment.

Transactions should be captured in Azure SQL Database.



Preferred solution

Commerce

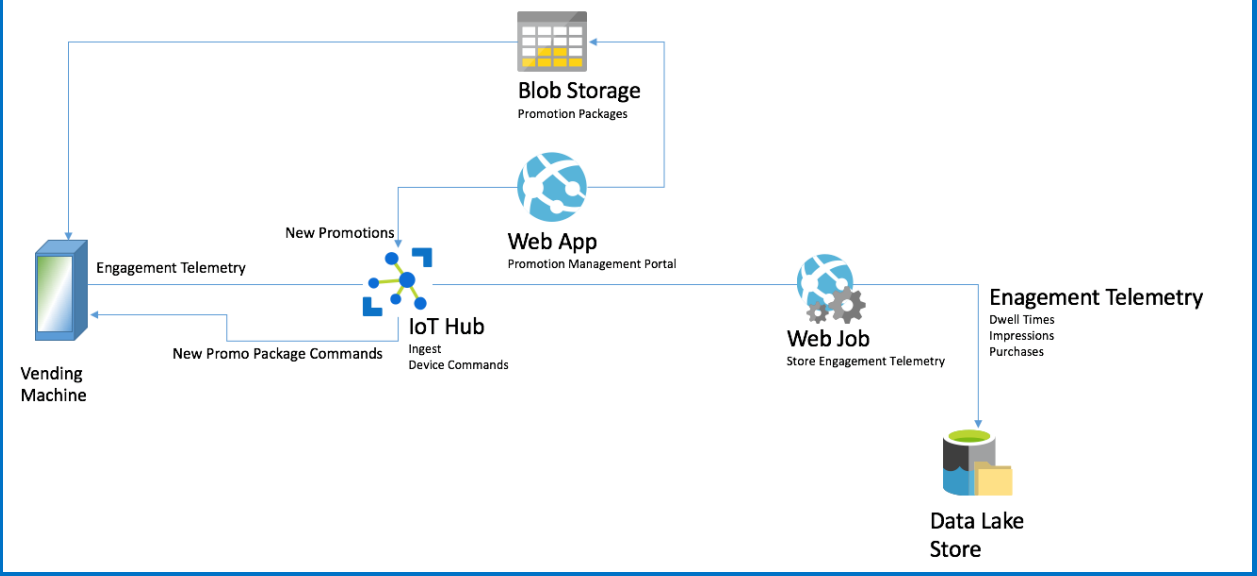
What services would you suggest and how would configure the storage and indexing?

To support the write-heavy workload that results from storing successful and aborted transactions, Trey should configure the transactions table as an In-Memory table with a Non-Clustered Columnar Index.

Durability of SCHEMA_AND_DATA, enabling the benefit of supporting large numbers of insert queries without the risk of losing data because it provides fully durable transactions that write the transaction log to disk before returning control to the client.

The Non-Clustered Columnar Index would be configured on all the columns. The effect of this is that the columnstore index maintains a copy of the data, enabling OLTP and analytics workloads to run against separate copies of the data. To leverage the memory-optimized features, Trey would need to deploy the Premium tier of Azure SQL Database.

Preferred solution



Engagement Analytics

What service would you recommend Trey capitalize on in order to scalably ingest the engagement telemetry directly from the vending machines?

Use IoT hubs to initially ingest and temporarily store engagement telemetry from the vending machines and to support intelligent promotions (since IoT Hub provides functionality for both device-to-cloud and cloud-to-device messaging).

Would you recommend they use Azure Storage Blobs or Azure Data Lake Store for persisting their engagement telemetry? Be specific with your reasoning.

Use Azure Data Lake Store as it addresses concern of "maxing out" their storage capacity.

Unlike Azure Storage Blobs (which has a maximum capacity of 500 TB per account), Azure Data Lake Store has no fixed upper limits and scales as storage needs scale. To accomplish a similar goal with Azure Storage would mean provisioning new Azure Storage accounts each time the limit is approached, and managing the storage of the data across multiple Azure Storage accounts.

Preferred solution

Engagement Analytics

What is processing the telemetry ingested, at least in so far as persisting the telemetry to the durable storage you recommended? How is this configured or implemented?

- The simplest option would be to configure an Azure Stream Analytics job to pull the data from IoT Hub and configure the Data Lake Store as the stream output
- Every 90 days the authorization for Stream Analytics to access Azure Data Lake Store needs to be renewed. To renew the authorization will require Trey to stop their Stream Analytics job, renew the authorization and restart it.
- Instead, use a Web Job that is running the Event Processor Host (from the Service Bus SDK), and implement an Event Processor that uses the Azure Data Lake Store SDK to write the telemetry pulled from IoT Hub.

Preferred solution

Facial Demographics

What Azure service or API would you suggest Trey utilize for determining demographics about a consumer from a photo of them taken by the vending machine? How is this provisioned and utilized?

Utilize the Face API, a part of Microsoft Cognitive Services.

The Face API exposes a Face Detect operation that, provided a URL to photo, returns demographics about the subject of the photo including features such as age, gender, facial expression, facial hair, head position and if any form of glasses (reading, sunglasses, swim goggles, etc.) are detected.

Provision a Cognitive Services account with an API type of Face API in the Portal.

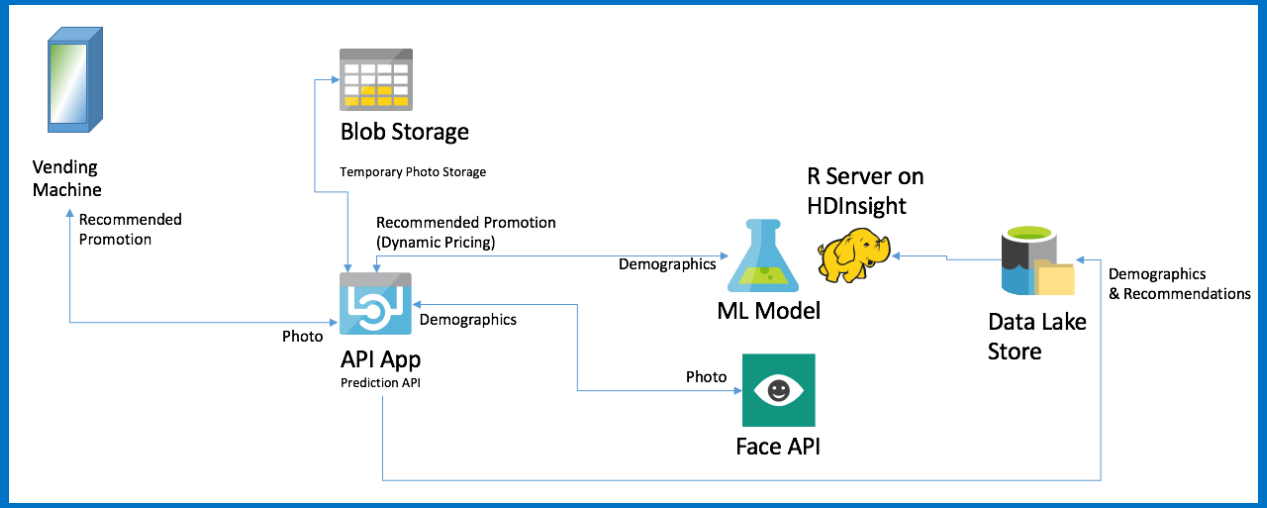
Provision an API App that would receive the photo from the vending machine, temporarily store it in Blob storage, and provide the URL (using a SAS token) to the Face API.

Once the demographics have been retrieved from the Face API, the API App can delete the photo from blob storage to preserve anonymity.

Preferred solution

Intelligent Promotions

What technology would you recommend that Trey use for implementing their machine learning model that recommends a product and price given consumer demographics?



First check that they really do need those large datasets to accurately train their model—in many cases very high accuracy can be achieved with datasets many times smaller, so it is important to involve a Data Scientist to verify this concern.

If large training data sets are required, Azure Machine Learning cannot be used (since the maximum dataset size it can access is 10GB).

Instead, Trey should look at running R Server on HDInsight. This would enable them to program their machine learning models in R, but also be able to address large training data sets as well as reducing training time by being able to parallelize the training across multiple nodes in a cluster.

Preferred solution

Intelligent Promotions

How would you guide Trey to load data so it can be used for training the machine learning model?

Trey should be capturing the telemetry to Azure Data Lake Store, which can be accessed from R Server on HDInsight.

What category of machine learning algorithm would you recommend to Trey for use in constructing their model? For this scenario your options are clustering, regression or two-class classification. Why?

Trey could implement their model using linear regression that would center around predicting the price at which to sell the items in the inventory of the vending machine, given the product and the demographics. A second model could be used that implements a two-class classification (sold or not sold) that predicts which product is likely to be bought given the product, demographics and recommended price.

Preferred solution

Intelligent Promotions

How would you operationalize your trained model so it can be invoked with the demographics?

They should expose the trained model as a web service.
One possible approach is to leverage Azure Machine Learning to host the trained model as a scoring service. To score by using an Azure Machine Learning web service, Trey could use the open source Azure Machine Learning R package to publish their model as an Azure web service. They would use the features of Azure Machine Learning to create an interface for the web service, and then call the web service as needed for scoring.

There is one caveat with this approach- Trey would need to convert any ScaleR model objects to equivalent open-source model objects for use with the web service. This can be done through the use of ScaleR coercion functions, such as as.randomForest() for ensemble-based models.

By doing this, Trey would be forgoing the scale out capabilities of the ScaleR objects.
To maintain the scale out capabilities, Trey could provision a server running Microsoft DeployR, which would enable them to deploy their R scripts and models as a web service. These web services can run models built with ScaleR functions and therefore, Trey would not need to sacrifice the scale-out training capability in order to achieve easy operationalization into web services.

Preferred solution

Intelligent Promotions

Where would you store the packages containing promotional artifacts for download by the vending machine and how would you instruct the vending machine to download and install them? Be specific on any Azure services used.

The packages could be stored in Blob storage. A SAS signature could be generated that the vending machine would use to download the package. This SAS URI would be provided by sending a command to the specific vending machine using IoT Hub.

Preferred solution

Visualization and reporting

What tool would you recommend Trey utilize for performing ad-hoc wrangling, exploration and visualization of their data?
They could use Power BI Desktop. The Query Editor functionality would enable them to filter and shape the data before visualizing it in reports and creating dashboards.

How would you make the resulting visualization available to others in the organization?
They can publish their reports to PowerBI.com.

Preferred objections handling

We've heard that Azure's machine learning can only train on data sets up to 10GB in size, are we blocked?

- While this is true for datasets processed using Azure Machine Learning, this is not the case in Azure when using R Server, either on HDInsight or as SQL Server R Services with SQL Server in a VM.

While not required in the short term, would our machine learning approach enable us to support reinforcement learning (whereby recommendations that lead to a purchase are preferred over time)?

- Reinforcement learning requires a training loop that feeds successful predictions back into the model to improve. Azure Machine Learning does not support this, but there are various R learners (that can run on R Server on HDInsight or SQL Server R Services) that support reinforcement.

Preferred objections handling

Can we really perform real-time analytics using only a single data store? We cannot afford to lose any of our purchase transaction data.

- Yes. Real-time analytics in SQL Server 2016 and Azure SQL Database is an approach that aims to provide support for analytics workloads (e.g., queries that compute aggregates over large portions of a dataset) while not affecting the performance or durability of transactional workloads (e.g., those that are write/query intensive). This is accomplished with the In-Memory and Columnar Index features.

We are concerned that training our models will take too long.

- Azure Machine Learning currently uses a single virtual machine instance to train a model and does not utilize algorithms that parallelize across cores. R Server on HDInsight provides specialized algorithms that parallelize across server cores and across nodes in the HDInsight cluster. SQL Server R Services does not currently parallelize training across servers in a cluster.

Customer quote

"We are predicting a future full of intelligent vending machines, thanks to Azure."

Grant Trey, Chief Innovation Officer, Trey Research