Streaming Machine Learning with Azure Databricks

Organizations are beginning to not only benefit from streaming data solutions, but require them to differentiate themselves from their competitors. Real-time reporting, alerts, and predictions are now common requests for businesses of all sizes.

That said, they rarely understand the requirements or implementation details needed to achieve that level of data processing. Streaming data is information that is generated and consumed continuously. This information typically includes many data sources, including log files, point of sale data (in store and online), financial data, and IoT Devices, to name just a few.

## Implementation

### **Fast and Easy**

Generating business-changing insights from streaming data can be a difficult process; however, there are quick wins for organizations of all sizes. Microsoft Azure offers an abundance of PaaS or SaaS products that allow users to connect to sources and automate workflows.

With Azure Logic Apps, it is extremely easy to set up data pipelines that extract data from your social media pages, analyze them for sentiment analysis, and alert users when comments or posts need to be addressed. While this may not be a business-changing solution, it gives companies the ability to have a more intimate level of interaction with customers or users than they had before.

Microsoft has provided a simple solution for companies to take advantage of this capability. Using Azure Logic Apps and Microsoft Cognitive Services, one can be alerted of any positive or negative tweets that occurs about their company. This is an easy and cost-effective way to implement intelligence into workflows. (Check out the example available [here](https://blogs.msdn.microsoft.com/deeperinsights/2017/07/12/how-to-measure-twitter-sentiment-with-azure-logic-apps-sql-database-and-power-bi/).) Azure Logic Apps connect to a variety of data sources, enabling organizations to obtain a quick win for real-time reporting with a deceptively simple drag-and-drop interface.

### **Ideal Implementation**

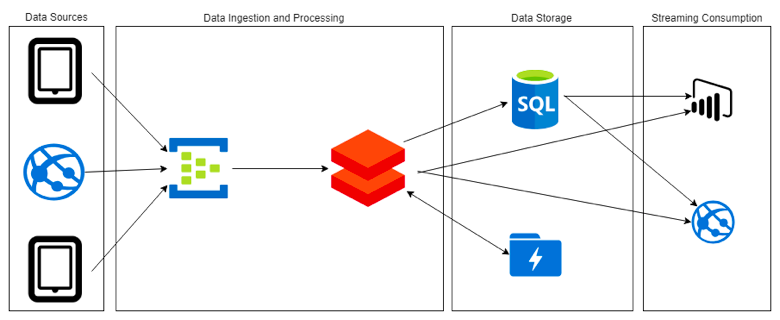
From my experience, companies benefit most from custom machine learning solutions that solve a specific business problem using their own data. Creating solutions tailored to solve a problem in a specific environment allows a business to truly take a proactive approach as they incorporate intelligence throughout their organization. However, lack of knowledge is often a barrier for companies when implementing custom and scalable solutions.

Azure Databricks is an optimized [Apache Spark](https://spark.apache.org/) platform perfect for data engineering and artificial intelligence solutions. It is an ideal platform for implementing batch or streaming processes on business critical data, and enables developers to create and deploy predictive analytics (machine learning and deep learning) solutions in an easy to use notebook environment.

Initially, organizations may implement their solutions as batch processes on Azure Databricks to save on cloud consumption costs, but plan for the future by using a platform that will scale and grow with the needs of the business. Batch processes allow users to save on monthly costs by turning off your virtual machines when they are not used, then when real-time insights is required the developer can almost flip a switch for streaming data. Deploy cost effective infrastructure now with the ability to scale limitlessly as you need in the future.

Below is a common infrastructure diagram I implement with my customers. If streaming is not required then we simply bypass the event hub and write python or scala scripts to connect directly to the data sources.

1. A number of data sources (devices, applications, databases etc.) that publish information to an Azure Event Hub (or [Apache Kafka](https://kafka.apache.org/)).
   * 1. Please note that whatever the data source is, there will always need to be some sort of process or application that collects data and sends it to the Event Hub.
2. Azure Databricks will write the stream of data as quickly as possible to an unaltered, "raw", data storage in an Azure Data Lake Store or Azure Blob Storage.
3. In addition to writing to raw storage, Databricks will be used to cleanse data as needed and stream appropriately to an application database, Power BI, or use [Databricks Delta](https://docs.databricks.com/delta/index.html) for real-time insights, consumption, and intelligent automated actions. Please note that applications can read directly off an Event Hub as a consumer as well.
4. Then use Azure Databricks to train a machine learning or deep learning model that can be used to make streaming or batch predictions.



### **Tips to Actually Implement a Solution**

When implementing new intelligent solutions with cloud infrastructure, it is likely that it will require internal business stakeholder buy in. Therefore, in order to successfully implement a new predictive analytics solution you must:

1.  Identify a business problem to solve and the stakeholders

2.  Visualize or surface results to "wow" stakeholders

3.  Start developing iteratively

If a team attempts to solve too many problems initially by trying to answer all possible questions, they will likely fail to "wow" a business user. Developers will likely focus all their time on coding and analyzing the best path forward that they will only have code to show (code is a rather boring deliverable for most business users), and may simply never get past the proof of concept or analysis phase.

#### Business Problem

It is common for companies to simply start creating a solution to work with newer technology without a true business problem they are trying to solve. It happens most often for organizations who want to start a data lake strategy. Their main goal is to develop a data lake so that other business units can take advantage of the sandbox environment for predictive analytics.

I believe a centralized data lake for organizations is a great idea for any IT group. However, without a specific business problem, it is difficult to see the true value that a data lake or machine learning solution provides, which in turn can slow adoption. By focusing on solving a single use case other, there will be a reference to other business units on *why* they should use the enterprise data lake. The reason for adoption is much more tangible.

#### Wow Stakeholders

There is not a more boring outcome to a business stakeholder than a project resulting in code. Machine learning or deep learning projects must have some type of end product that accurately describes the effectiveness of the solution created. In most machine learning solutions that I implement, I will almost always provide a [Power BI Report](https://powerbi.com). This ensures that the model and predictions are tangible because they are shown through visualizations. The business user now has the ability to actually use the predictions and show other internal users the solution.

#### Iterative Development

The most frustrating part of projects can be the initial planning or analysis phase. Large enterprises will often start a project and get stuck in analysis paralysis. I encourage teams I work with to simply start coding! This does not mean to do zero planning or proof of concepts, but at some point a team has to pick a direction and run with it. Avoid over analyzing various products by picking a small subset of well-known products, analyze them, and go!

## Benefits

Streaming data architecture is beneficial in most scenarios where dynamic data is generated on a continual basis. Any industry can benefit from data that’s available almost instantly from the time it was created. Most organizations will begin with simple solutions to collect log data, detect outliers based on set (unintelligent) rules, or provide real-time reporting.

However, these solutions evolve, becoming more sophisticated data processing pipelines that can learn and detect outlier data points as they occur. The true advantage of streaming data is in performing advanced tasks, like machine learning, to take preventive or proactive action.

Processing a data stream effectively generates quick insights, but it does not replace batch processes. Typically, organizations implement both solutions to gain quick, more computationally intensive insights. Streaming data reacts to or anticipates events, while batch processing derives additional insights after the fact.

Batch processing can often require more compute. It’s ideal when time or speed is not a priority. One of the biggest advantages of Azure Databricks is that companies are able to use the same infrastructure for both their workflows!

Batch processing data requires a system to allow data to build up so that it can be processed all at once. This often requires larger compute resources than streaming due to the size of data, which can be a hurdle for most organizations; however, it allows users to aggregate and analyze large amounts of data over a longer period of time. Streaming solutions do less computing, but require machines to be running 100% of the time and typically look at data over a shorter period of time.

## Example

I recently created a simple walkthrough of how to implement a streaming data solution on Azure. Check out the walkthrough on [GitHub](https://github.com/ryanchynoweth44/StreamingExampleDatabricks). Please note that an [Azure subscription](https://azure.microsoft.com/en-us) is required.

## Conclusion

Organizations of any size can benefit from a streaming solution using Databricks and Azure Data Lake Store. It enables near real-time reporting, as well as, provides a sandbox environment for iterative development of intelligent solutions. Azure Databricks and Data Lake Store allow a developer to implement both batch and streaming solutions in a familiar and easy to use environment.