The Accelerator for Apache Spark, Retail Use Case
Identify Patterns in Big Data

By: Hayden Schultz

TIBCO Accelerator for Apache Spark is an example of how TIBCO technologies and big data systems like Hadoop and Spark can be combined to perform offline analysis of historical patterns, consume streams of live data, and then act in real time when significant patterns are detected. The accelerator is a reusable set of components that handles extremely large volumes of data and solves big data problems in an extremely scalable and robust manner.

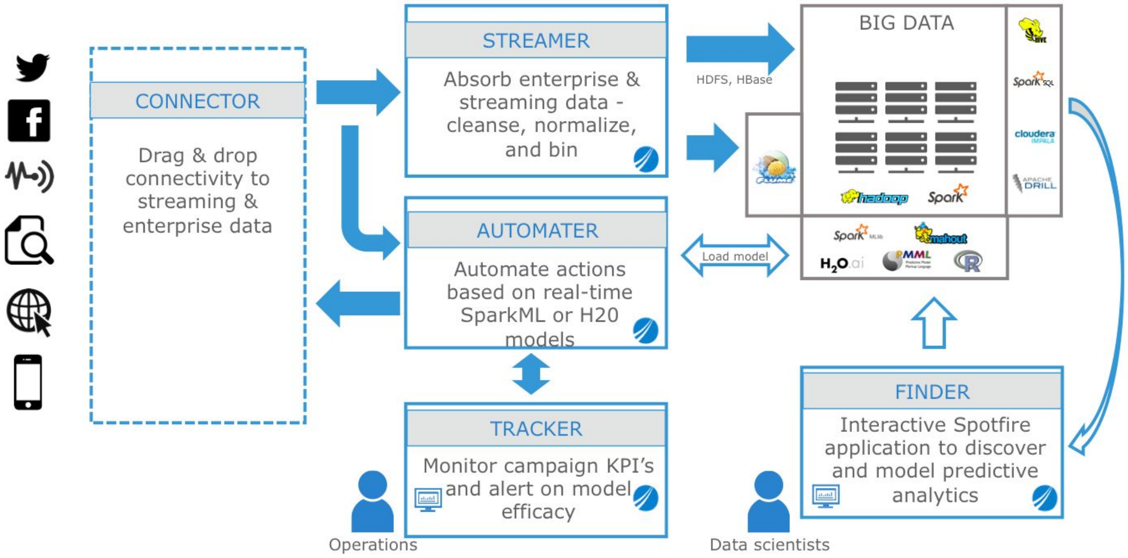
Traditionally, performing these analytics and event processing tasks requires many different development efforts. You need to write code to capture, store, and analyze data; train ad hoc, statistical, or machine learning models on the data; then connect to data streams and execute the trained models in real time. It's extremely common for these development efforts to use several programming languages and libraries so that the work achieved in each stage cannot be shared by developers in the other development stages. Development and analysis tools usually do not integrate with each other and the programs that were written to capture, train, or execute trained models.

The Accelerator for Apache Spark uses StreamBase, Spotfire, and TIBCO Enterprise Runtime for R (TERR) to capture and persist real-time data into a big data system, analyze the historical data in the big data system, train statistical and machine learning models in a big data cluster, and then execute the trained model without writing traditional code or sharing tools and separate development efforts. It is significantly faster to deploy big data applications using TIBCO tools that greatly speed the analysis, development, and testing of the entire application.

One use case is retail customers purchasing clothing through POS machines, online web stores, and other channels. Consumer purchasing trends change seasonally and vary as fashion trends also change.

A simple StreamBase application reads POS or web store data from an extension point such as a Kafka adapter, EMS/JMS adapter, or any other input source. The customer information is matched for the purchase event, then the customer segment is looked up. The purchase, customer, and customer segmentation information is passed into a propensity model that makes an upsell recommendation that is sent back to the data source.

The data is normalized and stored in the big data system. A large variety of formats can be selected, including hdfs csv files, hdfs binary files, HBase, Avro, or Parquet.


CONNECTOR
Drag & drop connectivity to streaming & enterprise data
STREAMER
Absorb enterprise & streaming data - cleanse, normalize, and bin
AUTOMATER
Automate actions based on real-time SparkML or H20 models
TRACKER
Monitor campaign KPI's and alert on model efficacy
Operations
HDFS, HBase
BIG DATA
Load model
FINDER
Interactive Spotfire application to discover and model predictive analytics
Data scientists

The retailer can use Spotfire to load customer and sales data stored in the big data cluster to create customer segmentation and sales propensity models. Spotfire allows the user to inspect the data, transform it, choose statistical or machine learning models, choose a modeling system like Spark ML, H2O, or arbitrary R code, and then start a training job on the big data cluster using Spark, MapReduce, or TERR. The trained models are then serialized so they can be executed against real-time data in StreamBase.

As purchases and upsell recommendations are issued, the application also keeps track of the success rate of the recommendations. Seasonal changes and fashion trends will change the products that are the most likely to be purchased by different customer segments, so the customer propensity models will loose effectiveness with time. When the success rate for the customer propensity segmentation model falls below a threshold, the StreamBase application will start a new job on the big data cluster to train a new model using the date when the old model's performance began to decline. When the new model has been trained, the StreamBase application will load it, replacing the old version.

Each of the input and modeling components in the StreamBase application are done with extension points. This allows different input sources or different modeling algorithms to be swapped in without changing the rest of the application.

In addition to this specific retail use case, you can see how the TIBCO Accelerator for Apache Spark integrates TIBCO products with big data systems to provide reliable, highly scalable implementations.