



MLOps with Spark, MLflow and Azure

January 2020



Introductions



Raki Rahman

Solution Architect

slalom | Data & Analytics

<https://rakirahman.com>

Professional Experience

6+ years at Technology Consulting and Software Development firms:

- Azure Big Data Architecture
- Environment Operationalization: DevOps, MLOps, DataOps
- Data and Software Engineering
- Delivering and operationalizing Cloud Pipeline (Big Data, DevOps, ML) projects for some of the largest Enterprise Clientele in Canada
- Passionate about solving complex problems with elegance

Education

B.A.Sc. Aerospace Engineering (Engineering Science) – University of Toronto



Topic:

MLOps with Mlflow in Azure Databricks

00:00

slalom

Slalom Introduction

00:05

Project Showcase

 **Azure Databricks** + **Bloomberg**

Bloomberg overview and Project Goal

Azure Solution Architecture

Bitcoin POC Use case with Natural Language Processing

00:30

mlflow[™]

MLflow – End-to-end demo in Azure Databricks

00:60



strategy

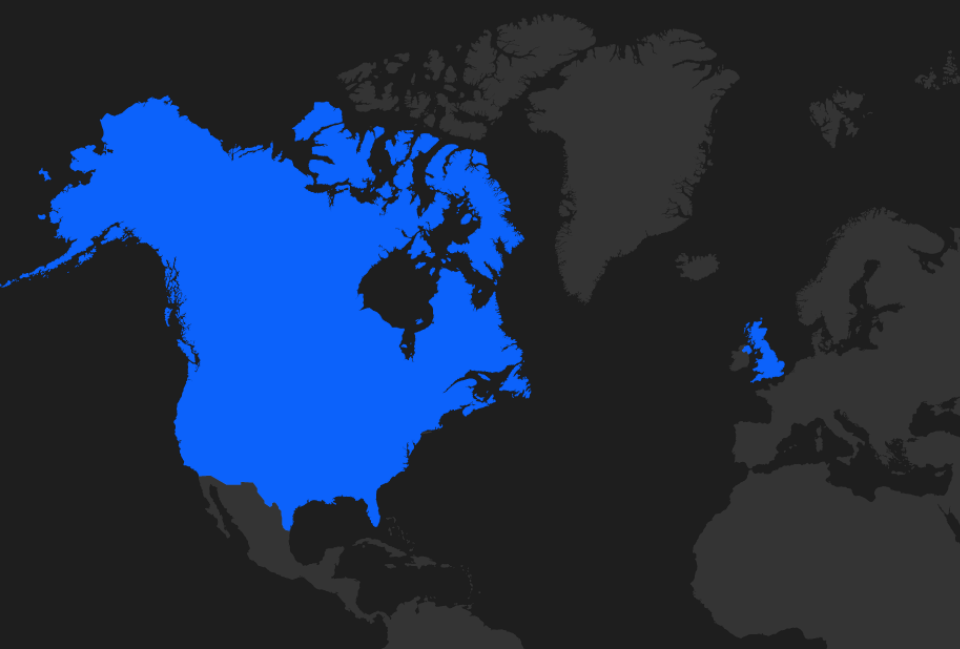
Redefine what is possible

technology

We analyze, architect, and co-create

transformation

We deliver with you



Atlanta • Austin • Boise • Boston • **Calgary**
Charlotte • Chicago • Dallas • Denver • Detroit
East Bay • Florida • Fort Worth • Hartford
Houston • London • Los Angeles • Manchester
Minneapolis • **Montreal** • New York • Orange
County Philadelphia • Phoenix • Portland • Salt Lake
City San Diego • San Francisco • Seattle • Silicon
Valley St. Louis • **Toronto** • **Vancouver** •
Washington

Slalom Data & Analytics (D&A)

Slalom is a modern consulting firm focused on strategy, technology, and business transformation.

Love your future. Love your data. We help people improve the world with data-driven insights.

1500+

Data & Analytics Consultants

27

Markets with Experience

>20

% of Total Slalom

>500

Certifications

6

2018 Partner of the Year Awards

2

MIT/Sloan Analytics Hackathon Winner

What We Do



Data & Analytics Strategy



Modern Data Architecture



Data Management



Visual Analytics



Augmented Intelligence

We're honored to be recognized by our partners for our data and analytics work.

Microsoft

Gold Partner

Global Power BI Partner of the Year 2018

Communication & Collaboration Partner of the Year 2017

Databricks

Consulting & SI Partner

Snowflake

Partner award winner

5x Tableau

North American Partner of the Year

Alteryx

Partner of the Year 2018

Google Cloud

Premier Partner

AWS

Premier Consulting Partner

slalom

Our Point of View

The Slalom logo, consisting of the word "slalom" in a bold, blue, lowercase sans-serif font, is centered within a blue rounded square border.

Bloomberg overview



Bloomberg overview

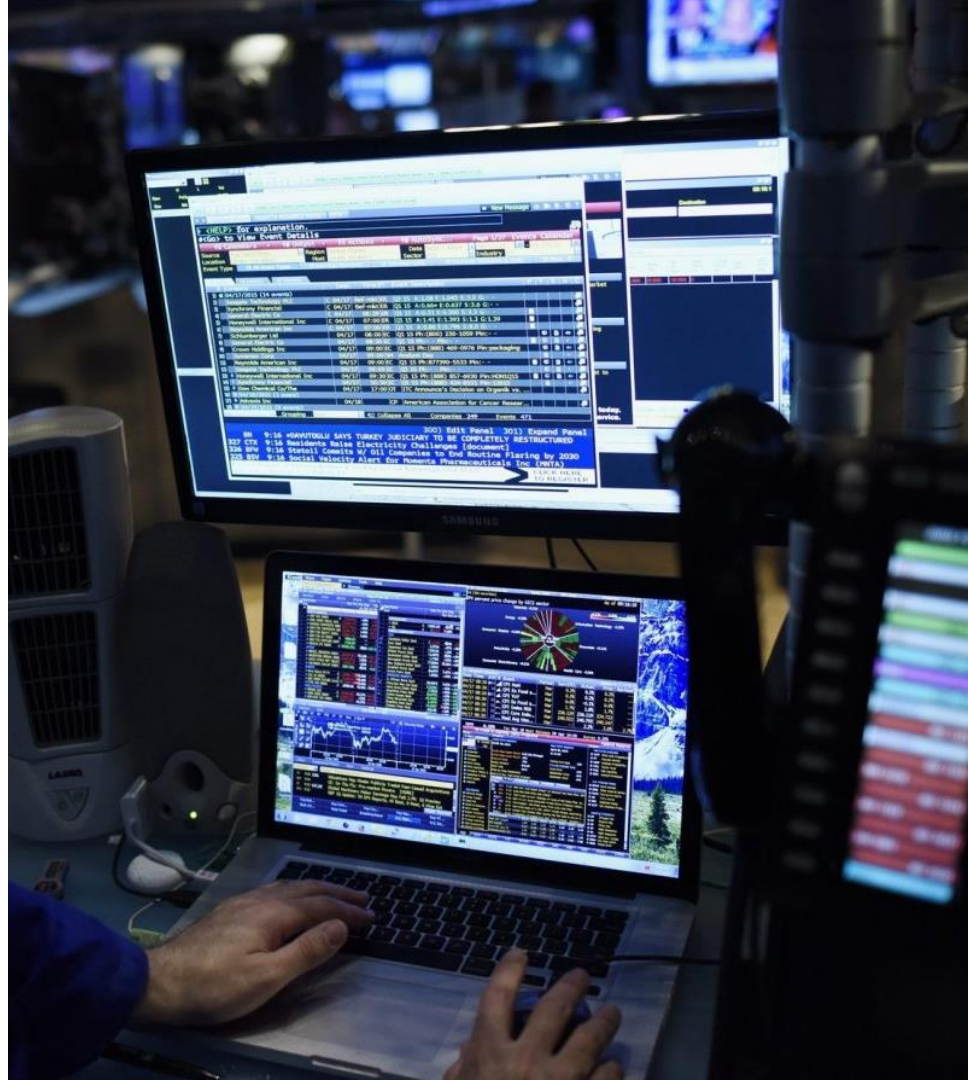
Massive global database of real-time and historical financial data – including trading, currency, supply chain, news, economic forecasts and more.

Bloomberg Terminal

The core revenue generator – implementing a client-side, physical terminal to connect to the Bloomberg Data Center.

Bloomberg Market Data Feed (B-PIPE)

B-PIPE enables real-time access to streaming data – allowing client applications access to the Bloomberg Data Center via the BLPAPI (C++, .NET, Java, Python).



Project Goal

Financial Services - Pension Fund Investment

Platform Requirements

**Architect a state-of-the-art, automated,
operationalized Cloud Data Analytics Platform**

Project Goal

Financial Services - Pension Fund Investment

Platform Requirements



Azure Databricks + Bloomberg

Real-Time Ingestion



Stream + Batch:

- Ticker data
- Share price
- Company news
- Supply chain updates
- Historical data

Secured API Integration



ExpressRoute
Integration to
Bloomberg API

Rapid ETL



High-speed Data
processing
(Stream + Batch
loads)

DevOps



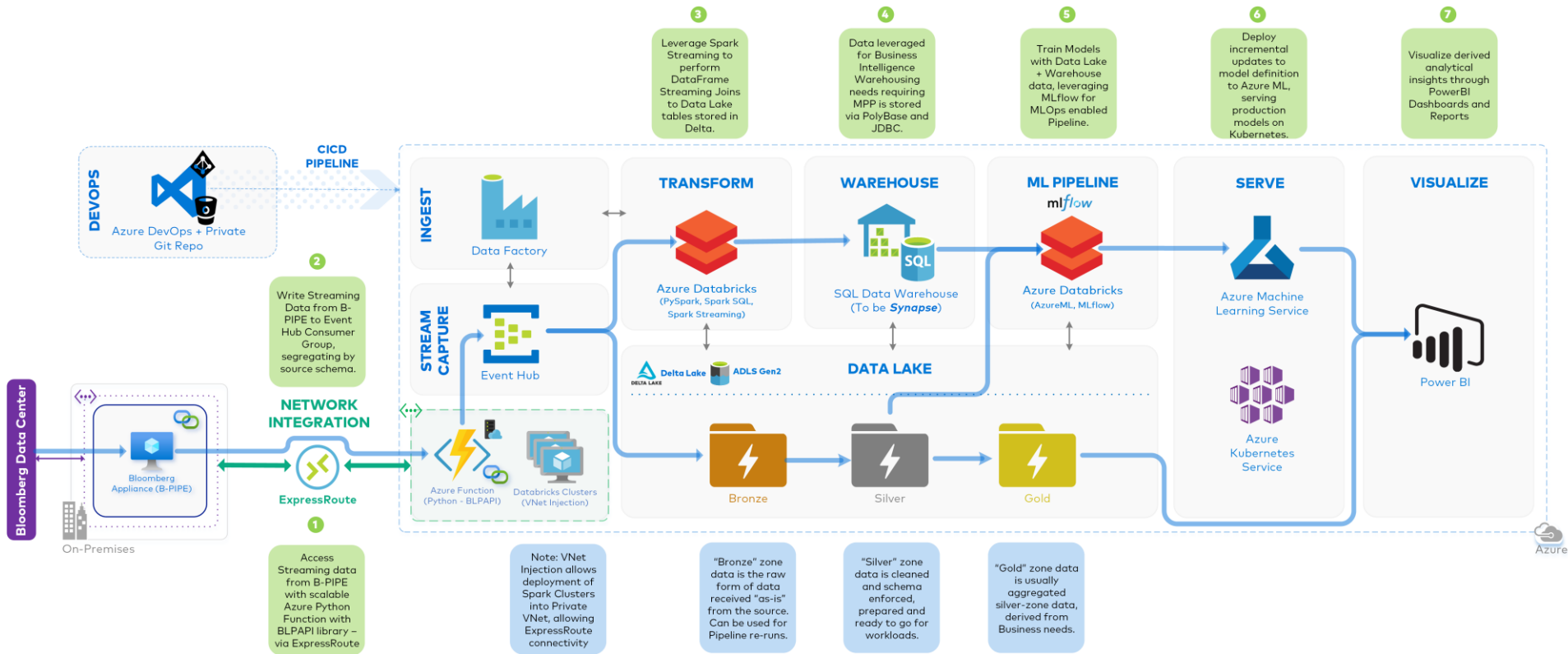
Continuous
Integration
Continuous Delivery
(CI/CD) enabled Data
Pipeline

Machine Learning Pipeline



MLOps enabled ML
Pipeline capable of
real time scoring and
incremental model
updates

Azure Solution Architecture



ML POC Use Case

Machine readable news with Bloomberg

Proof of Concept Objective

**Use Machine readable news from Bloomberg to explore
ML driven market trading strategy**

Machine readable news

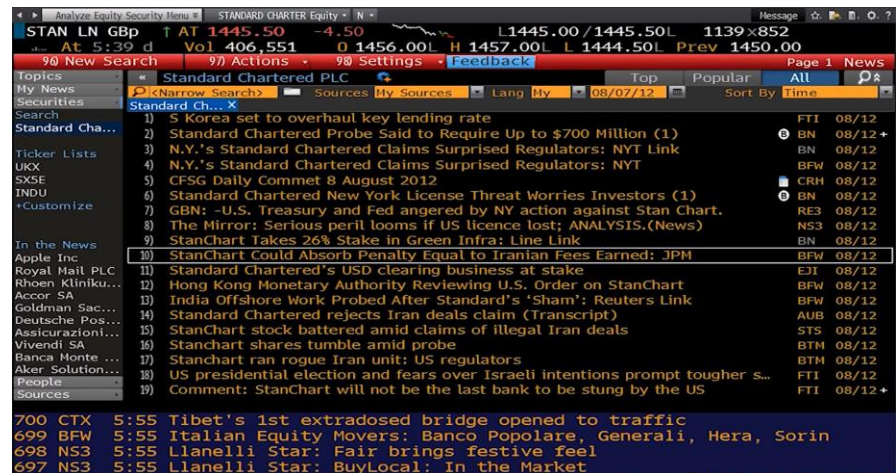
The idea is – Bloomberg provides access to real-time event-driven news feeds in the form of a structured dataset (i.e. consistently formatted, readership stats, sentiment, ticker tagged e.g. STAN LN).

The goal is – train NLP* driven Machine Learning models to filter news and generate Buy/Sell signals.

*NLP: training ML models to process and analyze large amounts of language data.


















On August 7th 2012, Standard Chartered stocks took a sharp decline.



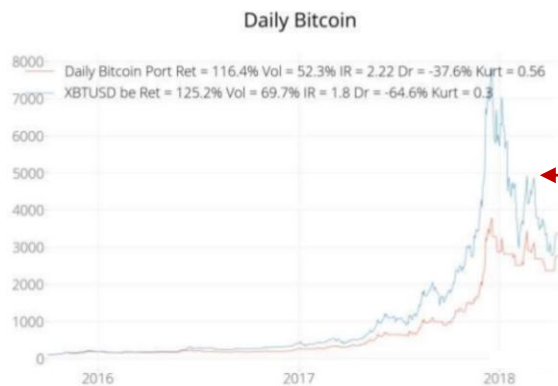
Viewing news from the timestamp shows root cause

Summarized high level approach

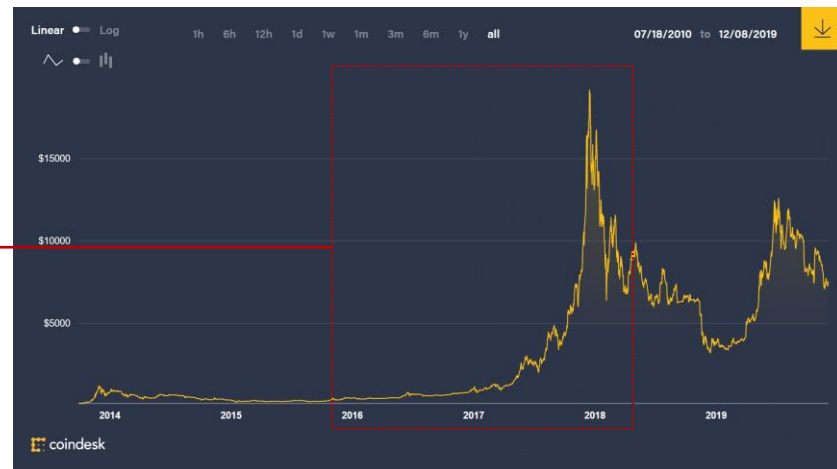
#	To do	Approach	Technology
1	Batch/Stream ingestion pipeline of ticker data (historical/real-time news + stock)	Pull historical/real-time data from B-PIPE into Data Lake and store in "Bronze" Zone. Create Data Ingestion pipeline for Stream and Batch processing.	 Azure Databricks  DELTA LAKE  Spark Streaming
2	Create Language Data Cleansing Pipeline for news data, and curate into time series format	Leverage Text Pre-Processing tools to perform cleaning operations (stemming, lemmatization, chunking/chinking) to filter noise and structure news data	 Koalas  BeautifulSoup  spaCy
3	Perform tokenization, topic modelling and sentiment analysis. Create indicators and apply trading rules.	Convert cleaned, structured text into mathematical (vectorized) representation, leveraging NLP specific mathematical techniques (TFIDF, TSNE, LSTM etc.). Convert into buy/sell signals and/or add other trading factors (e.g. carry)	 TensorFlow  Keras  Apache Spark ML
4	Score models on test data and deploy best performing model.	Perform experiments and track parameters/results using Mlflow. Deploy containerized models to Kubernetes using Azure ML version control.	 mlflow  Azure Machine Learning
5	Iterative Model Definition Updates	Connect B-PIPE Streaming pipeline to ML Pipeline for iterative scoring and incremental model updates.	 Azure Databricks  DELTA LAKE  mlflow  Azure Machine Learning

Trading Strategy Output from News Data

```
if __name__ == '__main__':  
    model = TradingBitCoinDaily()  
    model.construct_strategy()  
    model.plot_strategy_group_benchmark pnl()
```



Bitcoin (long) actual price VS.
Trading Strategy from NLP



Bitcoin price: 2010 to 2019

MLFlow end-to-end demo



An end-to-end demo showcasing MLflow **Tracking**, **Projects** and **Models**, using an Airbnb dataset.

We will finish by deploying a neural network model into Azure ML, and querying served model with Postman against the deployed Docker Container.

Link 1

<https://github.com/mdrakiburrahman/mlflow-end-to-end>

Link 2 (shortened)

<http://bit.ly/mlflow-demo>

The screenshot shows the GitHub repository page for `mdrakiburrahman / mlflow-end-to-end`. The repository is titled "MLflow end-to-end demo (tracking, projects, model) with Azure Databricks". It has 7 commits, 1 branch, 0 packages, 0 releases, and 1 contributor.

The README.md file is displayed, showing the title "mlflow-end-to-end" and the description "MLflow end-to-end demo (tracking, projects, model) with Azure Databricks".

Primary components

We're going to demo the three core components that make up mlflow:

- mlflow TRACKING**: Record and query experiments: code, data, config, results.
- mlflow PROJECTS**: Packaging format for reproducible runs on any platform.
- mlflow MODELS**: General model format that supports diverse deployment tools.

Links to the components are provided: databricks.com/mlflow, mlflow.org, github.com/mlflow, and twitter.com/MLflow.

Note: you don't have to use all three, each feature can be used independently.

Tracking

This allows us to log all aspects of the ML process - like different hyperparameters we tried, evaluation metrics, as well as the code we ran - alongside other arbitrary artifacts such as test data.

This also provides a *leaderboard-style UI* that makes it easy to see which model performed the best.

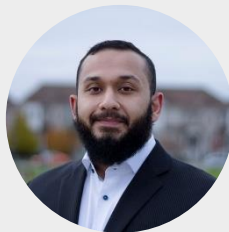
Projects

These are all about reproducibility and sharing. They combine *GIT*, the environment/model framework, either *conda* or *docker* and the specification that makes the code re-runnable.

Models

An abstraction that allows us to create/export models from any open source framework via the *Tracking* and *Projects* abstractions. We can also export them to a standard format that can be deployed to any number of systems. Since most deployment systems use some sort of container based solution (e.g. *AzureML* or *Sagemaker*), models make easy deployments to these systems - or we can deploy directly to *Kubernetes* or *Azure Container Registry*.

Thank you!



Raki Rahman

Data & Analytics
Modern Data Architect

raki.rahman@slalom.com