Streaming Outlier Analysis for Fun and Scalability

Casey Stella

2016

Table of Contents

Introduction

Hi, I'm Casey Stella!

Streaming Analytics

- The future involves non-trivial analytics done on streaming data
- It's not just IoT
- There is a need for insights to keep pace with the velocity of your data

Streaming Analytics

- **The Good:** Much of the data can be coerced into timeseries

Streaming Analytics

- **The Good:** Much of the data can be coerced into timeseries
- **The Bad:** There is a lot of data and it comes at you fast

Streaming Analytics

- **The Good:** Much of the data can be coerced into timeseries
- **The Bad:** There is a lot of data and it comes at you fast
- **The Good:** Outlier analysis or anomaly detection is a killer-app

Streaming Analytics

- **The Good:** Much of the data can be coerced into timeseries
- **The Bad:** There is a lot of data and it comes at you fast
- **The Good:** Outlier analysis or anomaly detection is a killer-app
- **The Bad:** Outlier analysis can be computationally intensive

Streaming Analytics

- **The Good:** Much of the data can be coerced into timeseries
- **The Bad:** There is a lot of data and it comes at you fast
- **The Good:** Outlier analysis or anomaly detection is a killer-app
- **The Bad:** Outlier analysis can be computationally intensive
- **The Good:** There is no shortage of computational frameworks to handle streaming

Streaming Analytics

- **The Good:** Much of the data can be coerced into timeseries
- **The Bad:** There is a lot of data and it comes at you fast
- **The Good:** Outlier analysis or anomaly detection is a killer-app
- **The Bad:** Outlier analysis can be computationally intensive
- **The Good:** There is no shortage of computational frameworks to handle streaming
- **The Bad:** There are not an overabundance of high-quality outlier analysis frameworks

Outlier Analysis

Outlier analysis or anomaly detection is the analytical technique by which "interesting" points are differentiated from "normal" points. Often "interesting" implies some sort of error or state which should be researched further.

[1] http://arxiv.org/pdf/1603.00567v1.pdf

Outlier Analysis

Outlier analysis or anomaly detection is the analytical technique by which "interesting" points are differentiated from "normal" points. Often "interesting" implies some sort of error or state which should be researched further.

Macrobase[1], an outlier analysis system built for IoT by MIT and Stanford and Cambridge Mobile Telematics, noted several properties of IoT data:

- Data produced by IoT applications often have come from some "ordinary" distribution
- IoT anomalies are often systemic
- They are often fairly rare

[1] http://arxiv.org/pdf/1603.00567v1.pdf

Outlier Analysis: A Hybrid Approach

In order to function at scale, a two-phase approach is taken

- For every data point

Outlier Analysis: A Hybrid Approach

In order to function at scale, a two-phase approach is taken

- For every data point
  - Detect outlier candidates using a robust estimator of variability (e.g. median absolute deviation) that uses distributional sketching (e.g. Q-trees)
  - Gather a biased sample (biased by recency)

Outlier Analysis: A Hybrid Approach

In order to function at scale, a two-phase approach is taken

- For every data point
  - Detect outlier candidates using a robust estimator of variability (e.g. median absolute deviation) that uses distributional sketching (e.g. Q-trees)
  - Gather a biased sample (biased by recency)
  - **Extremely deterministic in space and cheap in computation**

Outlier Analysis: A Hybrid Approach

In order to function at scale, a two-phase approach is taken

- For every data point
  - Detect outlier candidates using a robust estimator of variability (e.g. median absolute deviation) that uses distributional sketching (e.g. Q-trees)
  - Gather a biased sample (biased by recency)
  - **Extremely deterministic in space and cheap in computation**
- For every outlier candidate
  - Use traditional, more computationally complex approaches to outlier analysis (e.g. Robust PCA) on the biased sample

Outlier Analysis: A Hybrid Approach

In order to function at scale, a two-phase approach is taken

- For every data point
  - Detect outlier candidates using a robust estimator of variability (e.g. median absolute deviation) that uses distributional sketching (e.g. Q-trees)
  - Gather a biased sample (biased by recency)
  - **Extremely deterministic in space and cheap in computation**
- For every outlier candidate
  - Use traditional, more computationally complex approaches to outlier analysis (e.g. Robust PCA) on the biased sample
  - **Expensive computationally, but run infrequently**

Outlier Analysis: A Hybrid Approach

In order to function at scale, a two-phase approach is taken

- For every data point
  - Detect outlier candidates using a robust estimator of variability (e.g. median absolute deviation) that uses distributional sketching (e.g. Q-trees)
  - Gather a biased sample (biased by recency)
  - **Extremely deterministic in space and cheap in computation**
- For every outlier candidate
  - Use traditional, more computationally complex approaches to outlier analysis (e.g. Robust PCA) on the biased sample
  - **Expensive computationally, but run infrequently**

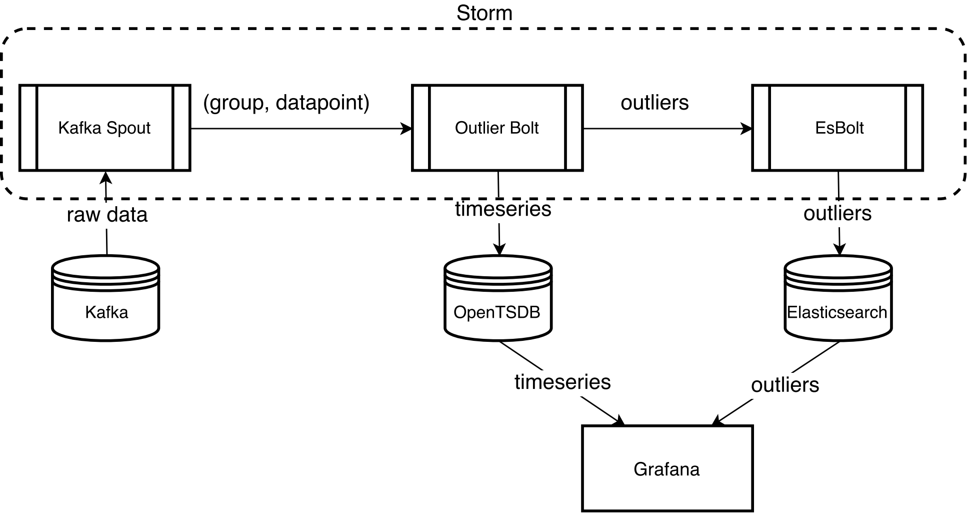
This becomes a data filter which can be attached to a timeseries data stream within a distributed computational framework (i.e. Storm, Spark, Flink, NiFi) to detect outliers.

Example: Streaming Financial Audit

Sometimes, doctors and hospitals have financial relationships with health care manufacturing companies. These relationships can include money for research activities, gifts, speaking fees, meals, or travel. The Social Security Act requires CMS to collect information from applicable manufacturers and group purchasing organizations (GPOs) in order to report information about their financial relationships with physicians and hospitals.

Let's treat each type (e.g. gifts, travel expenses) and physician specialty as a timeseries and look for anomalous payments.

Example Architecture



Questions

Thanks for your attention! Questions?

- Code & scripts for this talk available at http://github.com/cestella/streaming_outliers
- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cstella@hortonworks.com