Advanced Analytics

High-throughput, Scalable Data Stream Clustering

MARKET NEED

High volume, high throughput data streams are common in many industries including financial services (transaction streams), communications (instant messaging, SMS, micro-blogging), web and gaming (action and event streams) and production line environments (machine generated data). The ability to analyse and gain insights from this type of data as events happen, in real-time, can be hugely beneficial.

Traditionally data analytics is performed as an off-line, batch process where the results are available hours or even days after the data was produced. This means that any actions taken based on these insights will be at a considerable time interval after the original events occurred, and in many scenarios being able to analyse the live data stream and hence reduce this response delay is of critical importance.

Clustering is a core data analytics technique whereby similar entities are automatically identified and grouped together. This drives many common applications of data analytics such as detecting anomalous or fraudulent activity, identifyting market segments and user behaviours, reporting spam and emerging topics and patterns. CeADAR has developed a high-throughput, scalable clustering solution for data streams that brings real-time, 'live data' capabilities to these advanced data analytics tasks.

TECHNOLOGY SOLUTION

CeADAR's high-throughput, continuous clustering solution can process over 3 million entities per minute (50,000+ per second). Clusters of similar entities are automatically identified in the data stream and reported, along with associated statistics such as cluster size and growth rate, in near real-time. Although the system has been initially evaluated on textual data, the solution can be adapted to other content types such as transactions, images and other more complex data objects.

The technology is implemented on Storm, the open source data stream processing framework used by big data companies such as Twitter, Yahoo, Groupon and Klout. Storm enables scalability and the ability to run over commodity hardware or in the cloud – as data volumes grow or shrink, new servers or cloud instances can be easily provisioned to match requirements.

The continuous clustering technology uses a parallel clustering algorithm developed by CeADAR which harnesses LSH (Locality Sensitive Hashing), a technique that allows the parallel processing of the clustering task across multiple computing nodes. This allows the clustering to be applied on high volume, high throughput data streams.

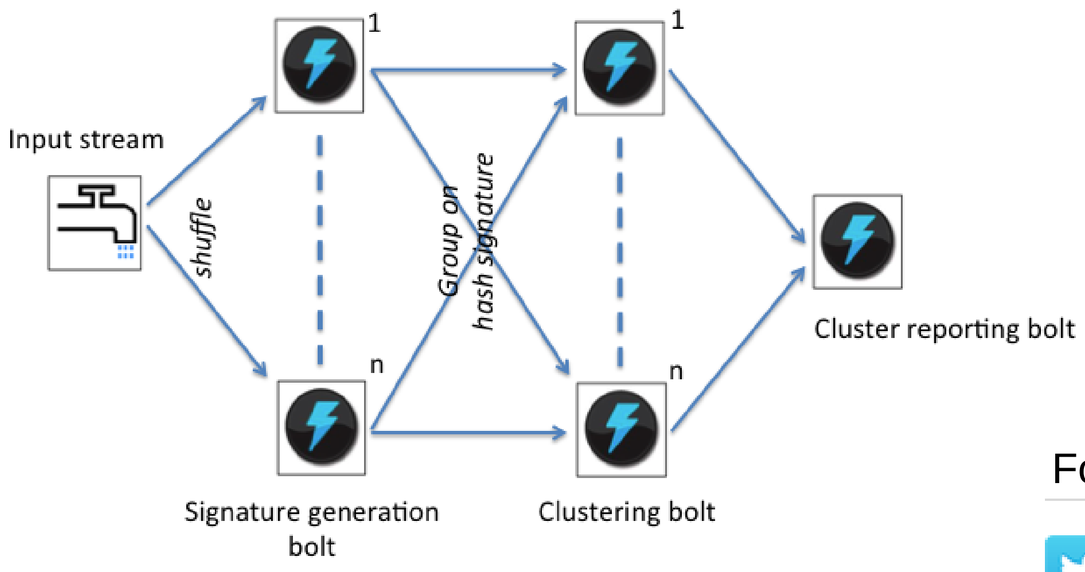

Figure 1: High level overview of Storm implementation

APPLICABILITY

Clustering, the automatic identification and grouping of similar items, is a fundamental technique in many high-level data analytics tasks including:

- Detecting and reporting new user behaviours and patterns
- Identifying fraudulent and anomalous behaviours
- Detecting spam in communications networks
- Reporting emerging topics and themes in content streams
- Understanding market segments and user types

There are many other domain-specific and niche tasks in which clustering can also be applied.

CeADAR's continuous clustering technology can be applied to high throughput, high volume data streams and enables these analytics tasks to be carried out live on the data. Although the initial focus has been on textual data, the core solution is data agnostic and can be adapted for clustering other data types including transactions, user actions, images etc.

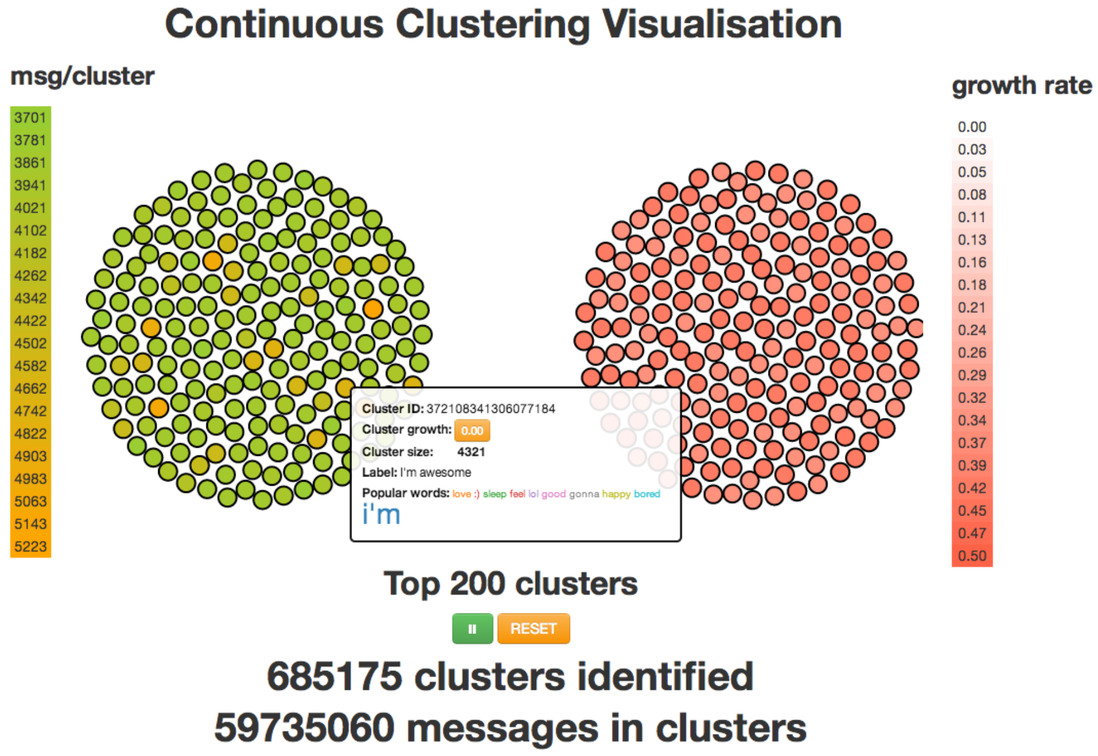

Figure 2: Interactive animated visualization of continuous clustering of the live Twitter stream showing top 200 largest and top 200 fastest growing clusters

RESEARCH TEAM

Dr. Oisín Boydell, UCD School of Computer Science and Informatics

Prof. Pádraig Cunningham, UCD School of Computer Science and Informatics

Dr. Marek Landowski, UCD School of Computer Science and Informatics

Dr. Guangyu WuUCD School of Computer Science and Informatics

Follow **CeADAR Ireland**:

Data Analytics

AN ENTERPRISE IRELAND & IDA IRELAND INITIATIVE