Real Time Analytics: Algorithms and Systems

Arun Kejariwal
Machine Zone Inc.
arun_kejariwal@acm.org

Sanjeev Kulkarni, Karthik Ramasamy
Twitter Inc.
{skulkarni, kramasamy}@twitter.com

ABSTRACT

Velocity is one of the **4 Vs** commonly used to characterize Big Data [5]. In this regard, Forrester remarked the following in Q3 2014 [8]: *"The high velocity, white-water flow of data from innumerable real-time data sources such as market data, Internet of Things, mobile, sensors, click-stream, and even transactions remain largely unnavigated by most firms. The opportunity to leverage streaming analytics has never been greater."* Example use cases of streaming analytics include, but not limited to: (a) visualization of business metrics in real-time (b) facilitating highly personalized experiences (c) expediting response during emergencies. Streaming analytics is extensively used in a wide variety of domains such as healthcare, e-commerce, financial services, telecommunications, energy and utilities, manufacturing, government and transportation.

In this tutorial, we shall present an in-depth overview of streaming analytics – applications, algorithms and platforms – landscape. We shall walk through how the field has evolved over the last decade and then discuss the current challenges – the impact of the other three **V**s, viz., **V**olume, **V**ariety and **V**eracity, on Big Data streaming analytics. The tutorial is intended for both researchers and practitioners in the industry. We shall also present state-of-the-affairs of streaming analytics at Twitter.

1. INTRODUCTION

Big Data is characterized by the increasing volume (of the order of zetabytes), and the velocity of data generation [6, 9]. It is projected that the market size of Big Data will climb up from the current market size of $5.1 billion to $53.7 billion by 2017 [1]. In recent years, Big Data analytics has been transitioning from being predominantly offline (or batch) to primarily online (or streaming). The trend is expected to become mainstream owing to the various facets, exemplified below, of the emerging data-driven society [10].

- Social media: Over 500M tweets are created everyday. A key challenge in this regard is how to surface the most personalized content in real time.
- Internet of Things (IoT): By 2020, the number of connected devices is expected to grow by 50% to 30 billion [4]. Data from embedded systems - the sensors and systems that monitor the physical universe - is expected to rise to 10% (from the current 2%) of the digital universe by 2020.
- Health Care: Increasingly Big Data is being leveraged in health care to, for example, improve both quality and efficiency in health care areas such as readmissions, adverse events, treatment optimization, and early identification of worsening health states or highest-need populations [11]. The volume of healthcare data is expected to swell to 2,314 exabytes by 2020, from 153 exabytes in 2013 [7].
- Machine data: With cloud computing becoming ubiquitous, machine generated data is expected to grow to 40% of the digital universe by 2020 [3].
- Connected vehicles: New telematics systems and the installation of ever greater numbers of computer chips, applications, electronic components and many other components provide data on vehicle usage, wear and tear, or defects [2]. The volume of data transferred per vehicle per month is expected to grow from around 4 MB to 5 GB. Further, by 2016 as many as 80% of all vehicles sold worldwide are expected to be "connected".

Elements of a data stream need to be processed in real time, else one may lose the opportunity to process them at all. Thus, it is critical that the data footprint of the algorithm fits in the main memory. Also, in light of the real-time constraint, it may be preferable to compute an approximate solution than an exact solution. Research in approximation algorithms for problems defined over data streams has led to some general techniques for data reduction and synopsis construction, including: *sampling, sliding windows, clustering, sketches, histograms* and *wavelets*.

Further, to be able to support Web scale and high velocity data, the algorithms should intrinsically distribute computation across multiple nodes and, if required, across data centers. In other words, the algorithms should be able to scale out.

In light of the dynamic nature of streaming data, a field of incremental machine learning has emerged to cater to Big Data streaming analytics. The techniques being developed are designed to work with incomplete data, to identify hidden variables to help steer future data collection and to quantify the change between one or more states of the model. Table 1 lists some of the most common problems addressed in prior research in the domain of streaming analytics and their example applications in the real world.

In early 2000s, Stream Processing Engines (SPEs) such as Aurora, STREAM, TelegraphCQ and Borealis were proposed. However, these systems did not scale with the increasing velocity and volume of the data streams characteristic of modern systems. To this end, several streaming platforms have been developed in the industry. Examples include, S4, Samza, Sonora, Millwheel, Photon, Storm, Flink,

Proceedings of the VLDB Endowment, Vol. 8, No. 12

Problem	Description	Application
Sampling	Obtain a representative set of the stream	A/B Testing
Filtering	Extract elements which meet a certain criterion	Set membership
Correlation	Find data subsets (subgraphs) in (graph) data stream which are highly correlated to a given data set	Fraud detection
Estimating Cardinality	Estimate the number of distinct elements	Site audience analysis
Estimating Quantiles	Estimate quantiles of a data stream with small amount of memory	Network analysis
Estimating Moments	Estimating distribution of frequencies of different elements	Databases
Finding Frequent Elements	Identify items in a multiset with frequency more than a threshold θ	Trending Hashtags
Counting Inversions	Estimate number of inversions	Measure sortedness
Finding Subsequences	Find Longest Increasing Subsequences (LIS), Longest Common Subsequence (LCS), subsequences similar to a given query sequence	Traffic analysis
Path Analysis	Determine whether there exists a path of length $\leq \ell$ between two nodes in a dynamic graph	Web graph analysis
Anomaly Detection	Detect anomalies in a data stream	Sensor networks
Temporal Pattern Analysis	Detect patterns in a data stream	Traffic analysis
Data Prediction	Predict missing values in a data stream	Sensor data analysis
Clustering	Cluster a data stream	Medical imaging
Graph analysis	Extract unweighted and weighted matching, vertex cover, independent sets, spanners, subgraphs (sparsification) and random walks, computing min-cut	Web graph analysis
Basic Counting	Estimate $\hat{m}$ of the number m of 1-bits in the sliding window (of size n) such that $\lvert\hat{m} - m\rvert \leq \epsilon m$	Popularity Analysis
Significant One Counting	Estimate $\hat{m}$ of the number m of 1-bits in the sliding window (of size n) such that if $m \geq \theta n$, then $\lvert\hat{m} - m\rvert \leq \epsilon m$	Traffic accounting

Table 1: Streaming algorithms and their applications

Platform	Description
S4	Real-time analytics with a key-value based programming model and support for scheduling/message passing and fault tolerance
Storm	The most popular and widely adopted real-time analytics platform developed at Twitter
Millwheel	Google's proprietary realtime analytics framework thats provides exact once semantics
Samza	Framework for topology-less real-time analytics that emphasizes sharing between groups
Akka	Toolkit for writing distributed, concurrent and fault tolerant applications
Spark	Does both offline and online analysis using the same code and same system
Flink	Fuses offline and online analysis using traditional RDBMS techniques
Pulsar	Does real-time analytics using SQL
Heron	Storm re-imagined with emphasis on higher scalability and better debuggability

Table 2: Open source streaming platforms

Spark, Pulsar and Heron. Some of these platforms have been open sourced. In order to be satisfy both, batch and streaming analytics, Lambda Architecture (LA) has been proposed as a robust, distributed platform to serve a variety of workloads, including low-latency high-reliability queries. Several platforms have been built based on the Lambda Architecture. Examples include Summingbird and Lambdoop. Commercial platforms such as TellApart are also based on the Lambda Architecture.

Table 2 summarizes a select streaming platforms developed over the years. In the tutorial, we shall walk the audience through the different design choices of the various platforms and the challenges which still remain. In addition, we shall also overview low-latency platforms built on top of Hadoop.

2. REFERENCES

[1] Big Data Market Size and Vendor Revenues. http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues.

[2] Connected cars get big data rolling. http://www.telekom.com/media/media-kits/179806.

[3] New Digital Universe Study Reveals Big Data Gap: Less Than 1% of Worlds Data is Analyzed; Less Than 20% is Protected. http://www.emc.com/about/news/press/2012/20121211-01.htm.

[4] The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. http://www.emc.com/leadership/digital-universe/2014iview/internet-of-things.htm.

[5] The Four V's of Big Data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

[6] Federal Government Big Data Rollout. http://www.nsf.gov/news/news_videos.jsp?cntn_id=123607&media_id=72174&org=NSF, 2012.

[7] K. Corbin. How CIOs Can Prepare for Healthcare 'Data Tsunami More like this. http://www.cio.com/article/2860072/healthcare/how-cios-can-prepare-for-healthcare-data-tsunami.html.

[8] M. Gualtieri and R. Curran. The Forrester Wave™: Big Data Streaming Analytics Platforms, Q3 2014. 2014.

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation, May 2011.

[10] A. S. Pentland. The data-driven society. *Scientific American*, 309:78–83, 2013.

[11] N. D. Shah and J. Pathak. Why Health Care May Finally Be Ready for Big Data. https://hbr.org/2014/12/why-health-care-may-finally-be-ready-for-big-data.