MacroBase: Prioritizing Attention in Fast Data

Stanford CS Future Data Systems Group
http://github.com/stanford-futuredata/macrobase
http://futuredata.stanford.edu/

MacroBase is a new kind of data analysis engine designed to *prioritize human attention* in large-scale *fast data* streams. Given increasing data volumes that are increasingly too large for manual inspection, MacroBase highlights interesting behaviors in streams and produces explanations for these behaviors that can be used for tasks including diagnostics, alerting, and root cause analysis. MacroBase has already found interesting, previously unknown behaviors and trends in production data in domains including mobile telematics, datacenter operations, electrical utilities, and satellite imaging. The purpose of this document is to describe the architecture, interaction model, and ongoing research in the MacroBase project.

Prioritizing Attention. MacroBase is designed to conserve the scarcest resource in fast data: human attention. Fast data streams exhibit challenges including a.) high volume (e.g., millions of events per minute), b.) heterogeneity (e.g., thousands to tens of thousands of device configurations; data source complexity including time-series, images, and video), and c.) demands for timeliness (e.g., minute- to second-granularity response time). As a result of these challenges, many important behaviors, including systemic inefficiency and unreliability, go unnoticed, resulting in degraded application performance, wasted resources, and unexpected failures. For example, the Android mobile operating system ecosystem currently includes over 24,000 distinct device types; given platform-specific differences in sensors, batteries, and processing capabilities, is a given mobile application operating correctly on each? In our experience with fast data from production deployments, pernicious behaviors lurk in the combinatorial explosion of hardware-firmware-software configurations. Systems should illuminate the combinations that matter.

MacroBase's Architectural Principles. MacroBase's design centers around two key tenets. First, MacroBase is designed for "pay as you go" functionality. MacroBase's default analysis pipeline provides results out of the box, without human labeling or need for domain knowledge. Subsequently, users can tailor results by encoding domain information, both by providing supervised feedback ("show me more/less like this") and by writing domain-specific feature extraction operators (e.g., image convolution). As a result, MacroBase is designed for ease of initial use, followed by improvements in result quality that are linearly (or super-linearly) proportional to the amount of end-user involvement. Because this task requires architectural extensibility, MacroBase's core operators are actually constructed using the same interfaces that end-user experts also use. Second, MacroBase is optimized to be lazy, to perform as little work as possible on each data point. In many applications, the most valuable information is contained in a small set of data; MacroBase aggressively prunes incoming streams to find this small set of data that matters. This pruning often takes the form of cross-layer optimizations, including density approximation and cardinality-aware summarization, that deliver many order of magnitude speedups in practice.

MacroBase's Operator Pipeline. Internally, MacroBase is powered by a combination of streaming classification and explanation operators and is the first engine to combine these techniques. By default, MacroBase finds unlikely events in data populations then searches for explanations that help differentiate statistically rare data points (e.g., "device 5012 is 48 times more likely to produce abnormally high sensor readings"). To accomplish this task, MacroBase adopts new, streaming density estimation algorithms to simultaneously and incrementally fit and classify data streams. Given labeled data or supervision rules, MacroBase executes a hybrid ensemble of supervised and unsupervised models, relying on the former to incorporate domain-specific knowledge and the latter to discover new, unknown behaviors. MacroBase incrementally produces explanations using new data structures including an improved efficient heavy-hitters sketch and a streaming prefix-tree-based index that enables fast search over interesting combinations. By configuring the feature transforms that feed this sequence of operators, MacroBase can analyze a range of diverse data, including time-varying sensor streams, video and images, and large warehouse-based datasets. MacroBase's output can feed both downstream reporting and alerting systems as well as automated remediation tools.

Ongoing Research. MacroBase is the vehicle for a number of ongoing research efforts including: feature extraction and fusion over heterogeneous data sources such as images, video, and sensor data; automatic optimization for streaming dimensionality reduction and data representation selection; scalable, online and incremental multi-modal density-based outlier detection; online ensembling of supervised and unsupervised detection methods; and new data summarization techniques for highlighting outliers in time-series and visual data. All of our research is driven by real problems encountered in real-world datasets and Internet of Things applications, often encountered in or gathered from production.

Our early experiences with attention prioritization have been highly encouraging. In contrast with the state of the art in many domains, which often constitutes statically-configured alerts and primarily manual root-cause analysis, MacroBase has highlighted new, interesting behaviors in production data. We are excited to continue exploring the power of MacroBase's "pay as you go" architecture to prioritize attention while further improving accuracy and performance.