While understanding and advocating Big Data is often within reach of most organizations, it is quite different when it comes to sharing real-world applications. It is often difficult to obtain access to large datasets that exhibit variety and quality. User experience needs to provide useful cues and perform interactions, for instance through advanced visualization. In order to experiment, master and demonstrate this value chain, the Computer Research Institute of Montreal (CRIM) started the Sherlock project in early 2014.

In order to build flexible capacities that can be adapted to an array of domains, our system relies on a modular approach that combines different technologies from the Hadoop ecosystem. The backend relies on Apache Spark for advanced analytics for Smart Data.

**Core Services and Platform**

Over the years, CRIM R&D teams had build an extensive portfolio of techniques and algorithms to process and enrich a large variety of unstructured data. Readily available components related to human activity are speech transcription, speaker identification, speech-to-text matching, keyword spotting and facial analysis. General purpose imaging components such as visual transition detection, feature extraction from satellite imaging and optical tracking are also part of the toolbox. These services provide structure to the documents, in the form of annotations, metadata or additional observations. Multimedia file storage and streaming, load balancing, batch processing, annotation storage and geospatial data management completes the bulk of the platform.

Implementations were recently deployed as reusables software components in the form of OpenStack virtual machines snapshots and Docker containers. CRIM is offering to its collaborators and partners experimental managed instances of these services. Gradually, analytic services and platforms will be hosted by dedicated managed hosting service providers.

**Use-cases**

The availability of speech, image, video and geospatial building blocks allows researchers, developers and deciders to address a wide array of use cases. Furthermore, CRIM’s strong NLP expertise has the potential to unlock even more value to data by adding semantic context. The platform is built with several use cases in mind, for instance video-processing, load balancing, batch processing, annotation storage and geospatial data management. These services provide structure to the documents, in the form of annotations, metadata or additional observations. Multimedia file storage and streaming, load balancing, batch processing, annotation storage and geospatial data management completes the bulk of the platform.

**Hardware**

The Sherlock Lab is hosted at CRIM and is backed by a cluster of high-performance servers with 480 CPU cores, 3.7 TB of RAM and 550 TB of storage. Servers are wired using a 10G network and connected to the CANARIE high-speed network.

**Research Results**

**Natural Language Processing**

Clustering, sometimes called unsupervised learning, is one of the most fundamental step in understanding a dataset, aiming to discover the unknown nature of data through the separation of a finite dataset, with little or no ground truth, into a finite and discrete set of “natural,” hidden data structures. Given a set of n points in a multidimensional space, the purpose of clustering is to group them into several sets based on similarity measures and distance vectors.

Most clustering algorithms have a quadratic complexity with n, making them unsuitable to analyze large amounts of data. In addition, many clustering algorithms are methods inherently difficult to parallelize.

We propose PatchWork, a novel clustering algorithm to address those issues. PatchWork is a mixture of density and grid-based clustering algorithm. It has linear complexity and near linear horizontal scalability. As a result, PatchWork can cluster a billion points in a few minutes only, a 100x improvement over Spark MLLib native implementation of the well-known K-Means.

**PatchWork clustering**

While understanding and advocating Big Data is often within reach of most organizations, it is quite different when it comes to sharing real-world applications. It is often difficult to obtain access to large datasets that exhibit variety and quality. User experience needs to provide useful cues and perform interactions, for instance through advanced visualization. In order to experiment, master and demonstrate this value chain, the Computer Research Institute of Montreal (CRIM) started the Sherlock project in early 2014.

In order to build flexible capacities that can be adapted to an array of domains, our system relies on a modular approach that combines different technologies from the Hadoop ecosystem. The backend relies on Apache Spark for advanced analytics for Smart Data.

**Core Services and Platform**

Over the years, CRIM R&D teams had build an extensive portfolio of techniques and algorithms to process and enrich a large variety of unstructured data. Readily available components related to human activity are speech transcription, speaker identification, speech-to-text matching, keyword spotting and facial analysis. General purpose imaging components such as visual transition detection, feature extraction from satellite imaging and optical tracking are also part of the toolbox. These services provide structure to the documents, in the form of annotations, metadata or additional observations. Multimedia file storage and streaming, load balancing, batch processing, annotation storage and geospatial data management completes the bulk of the platform.

Implementations were recently deployed as reusables software components in the form of OpenStack virtual machines snapshots and Docker containers. CRIM is offering to its collaborators and partners experimental managed instances of these services. Gradually, analytic services and platforms will be hosted by dedicated managed hosting service providers.