cloudera®

WHITE PAPER

Working With Images and Videos at Scale: Solutions for Law Enforcement

Version: 101

cloudera®

Table of Contents

cloudera

Breaking the Lock

Whether a consumer, a business, or a government organization, we use images and videos to capture and create a record of events about the world around us. We share that data to communicate in a manner that is often more efficient than the spoken or written word. But what exactly is the character of this data? A picture or a video frame is nothing more than an array of numbers arranged to give a color response at a specific location on a display screen (the red, blue, and green pixels we see on our monitors and TV sets). To understand the information displayed, our brains do a lot of work to interpret the visual appearance of those pixels and put them together in a way that makes sense of what is being displayed (objects, faces, text strings, scenes, etc.). In a very real sense, the data is locked in the image and the human brain is needed to unlock its meaning. While this might have been an acceptable approach in the past, it is not something that scales well. The sheer volume of media content being generated and the variety of use cases demand that we find other means of unlocking this value so that it can support large-scale applications and workflows. We need software that can do that work for us automatically and quickly; software that is adaptable to a range of applications and use cases; software that can meet the needs of big data in the enterprise.

Cloudera, piXlogic, and Intel have partnered to make advanced image and video analytics available to big data customers. The piXlogic software provides a scalable set of "extra eyes" that can automatically go through images and videos stored in Hadoop clusters and generate metadata describing their contents. This enables search, content discovery, alerting, and integration with other analytical tools in the Cloudera stack. Customer environments where images and videos are a critical component of the workflow will benefit greatly from these expanded capabilities. The Cloudera platform simplifies the management of distributed computing resources, such that piXserve can carry out its calculations where the data is stored. The platform provides for inherent horizontal scalability, and the system can grow as the amount of data grows. The level of analysis that piXserve carries out is very extensive and computationally expensive. Intel technologies are key to keeping computing costs low. This involves using both chips and accelerators with advanced extensions so that as much of the computations as possible are run on the bare metal for maximum performance.

In summary, the solution comprises three components integrated and working together:

- piXlogic brings much needed automation when dealing with images and videos;
- Cloudera brings an infrastructure platform that is proven at scale. Running the piXlogic software on Cloudera leverages these capabilities and allows horizontal scaling of the solution;
- Intel makes high performance computing possible on general purpose hardware. Running the piXlogic software on Intel hardware and accelerators enables economic deployments at scale.

The combination of the three creates a unique and compelling solution for law enforcement customers. These elements and the type of use cases that they enable are discussed subsequently.

The piXserve Component

piXserve is an award-winning, enterprise-class software application that helps customers automatically index, search, and tag media content. The software can process over 100 different image formats, including multi-gigapixel geospatial images. All major video formats are supported, as are live video sources from IP-addressable cameras or broadcast points. piXserve detects and recognizes visual objects, faces, and text strings that appear on the image/video frame. At its foundation, piXserve embodies very unique and strong image-analyses capabilities. Building on that are state-of-the-art machine learning and AI techniques. A unique aspect of the algorithms used in piXserve is that images/video frames are analyzed holistically, with image segmentation being a foundational component. This means that users can also search and work with things that were not recognized or deemed important enough to be specifically recognized or tagged at the time of indexing.

Several aspects make piXserve particularly compelling in workflows typically associated with law enforcement operations. The software operates on customer premises, so the data never leaves the customer data center, and the customer is in full control of every aspect. The software addresses and brings together in a single package many image/video-analytics features critical to the work

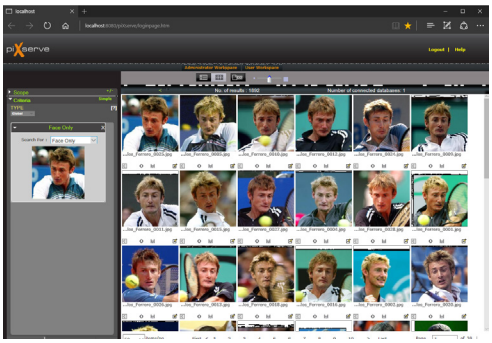


Fig. 1 State-of-the-art Face Recognition in piXserve

of analysts and field personnel. All this state-of-the-art capability can now be put into the hands of ordinary users, through a simple web interface (desktop or mobile). Working with images/videos need no longer be the realm of specialists. Workflows can be simplified and gains in productivity and speed can now be accrued more broadly.

Through a web browser interface, users can search and retrieve images and video segments that contain visual objects that are the same or similar to those in their query pictures. Users can also select specific items in their query image in order to search just for those. In addition, users can formulate search queries using more than one image as an input. In piXserve, a search query is encapsulated in a "search module," and multiple search modules can be combined to describe the full query, each module related to the others by means of standard logical operators such as "AND," "OR," and "NOT." Combined, these features provide an unparalleled level of flexibility and specificity to users (e.g., "find me images/video segments where the car in this picture AND the face in this other picture are both visible in the same shot").

piXserve can also read text that may appear in the field of view of the image, and do so in a variety of languages and character sets (Latin characters, Chinese ideograms, Korean, Japanese, Cyrillic, Hindu, Arabic, and many others). Users can type a text string and retrieve images/video sequences where such text appears in the image (for example, text on the screen of a news broadcast, text on a street sign, a license plate, etc.). piXserve can see and recognize text in complex color images and videos at a level of accuracy that is unmatched in the industry.

The piXserve visual search capabilities are complemented by extensive auto-tagging capabilities. When the software indexes visual media, it tries to add as many keywords/tags as it can to the records that are created to describe the contents of the images. piXserve creates keyword descriptors using several mechanisms, including:

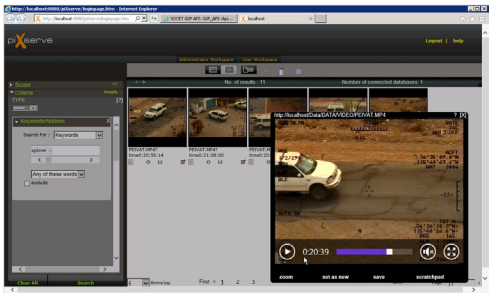


Fig. 2 State of the Art Object Recognition in piXserve

- **Face Naming** – Users can give a name to a person's face returned as a search result. When new images/videos are processed, if the detected face corresponds to a named face, that name is automatically assigned as a searchable keyword for that image/video frame.
- **Specific Object Recognition** – Users can give the software an image example of an item of interest together with the name of the item. During indexing, piXserve will check for the presence of that item and, if it sees it, capture its location on the image/video frame and assign the given name as a searchable keyword to that picture.
- **Generic Object Recognition** – piXserve can recognize certain generic classes of objects and concepts that may be visible in the image. When such items are detected/recognized, searchable keywords are automatically added to the records describing that video frame.
- **Keyword Recommendations** – piXserve can suggest keywords that could apply to an image/video frame based on images/videos with similar content that the customer has tagged in the past

Combined, these and many other features available in piXserve help enterprises automatically extract and work with information in photos and videos, and by leveraging the underlying Cloudera platform do so at scale.

The Cloudera Component

The right technology is key for turning data into real business value. Powered by Apache Hadoop, Cloudera Enterprise is the fastest, easiest, and most secure modern data platform. From analytics to data science, anyone can now get results from any data and across any environment—all within a single, scalable platform.

Cloudera Enterprise provides comprehensive data security and processing security, including authentication, authorization, fine-grained auditing, log redaction of sensitive information, encryption for data at rest and in motion, and encryption key management.

Authentication is provided by Kerberos for all internal machine-to-machine interactions. For end users, either Kerberos or LDAP is available for Cloudera Enterprise compute frameworks (MapReduce, Spark, Impala, and Search). Cloudera Enterprise supports using Active Directory for authentication (it is both an LDAP and Kerberos provider).

cloudera®

Authorization for record-based data is via Apache Sentry, which provides role-based access controls at record levels for Hive, Impala, Search (Solr), and any services that use HCatalog. Any files covered by HCatalog are automatically managed by Sentry. Sentry uses HDFS extended ACLs. Thus multiple groups/roles can be attached to a file's permissions for a given directory, file, or table. Sentry is a least-privilege model. If you can't see the whole table, then you can't read the file. All frameworks pass through Sentry or go straight to file. File access may be blocked due to Sentry settings. Cloudera is in beta for additional fine-grained controls at the HDFS level with RecordService (which supports non-relational interfaces Spark, Pig, MapReduce, etc.) Users get a filtered view of files based on their role(s).

Auditing is inescapable and performed at the framework and file levels. It is very fine-grained and detailed (see http://www.cloudera.com/documentation/enterprise/latest/topics/cn_iu_audits.html#cn_topic_7). Audit supports division of privilege so that auditors are neither Cloudera nor Linux administrators. Privileged commands such as those issued by Cloudera Manager are audited.

Cloudera audit logs and system logs can be configured to redact sensitive information (IP addresses, SSN, etc.). With redaction, sensitive information is never written to the logs. This is done via pattern matching and regular expressions.

Flexibility is a key characteristic of Cloudera Enterprise and what enabled us to work with piXlogic to help them integrate into the Hadoop environment. The underlying storage HDFS is a file system, so it can hold any type of data. The innovation in Hadoop is that multiple processing frameworks can read the same data and do different things with it. Hadoop was originally a batch processing environment. Cloudera led the way in creating new frameworks that could share the security and storage of Hadoop, but perform different actions on the data. Impala was created by Cloudera so that a high-performance interactive SQL engine could treat structured data as if it were an RDBMS. In keeping with our open source philosophy, it was donated to the Apache Software Foundation once it was feature rich and very stable. It is now Apache Impala. Cloudera Search (Apache Solr with some additional work to integrate it with HDFS) brought faceted text search to the platform. Cloudera was the first vendor to bring Apache Spark to the Hadoop ecosystem. Spark is a next-generation memory-oriented batch processing environment. It is often used not only for data transformation but also for machine learning.

Cloudera Enterprise and piXlogic are designed so they can be run on-premises or in the cloud. Cloudera Enterprise supports Amazon, Google, and Microsoft cloud environments. In the cloud, Cloudera Enterprise can be run either full time or ephemerally.

The Intel Component

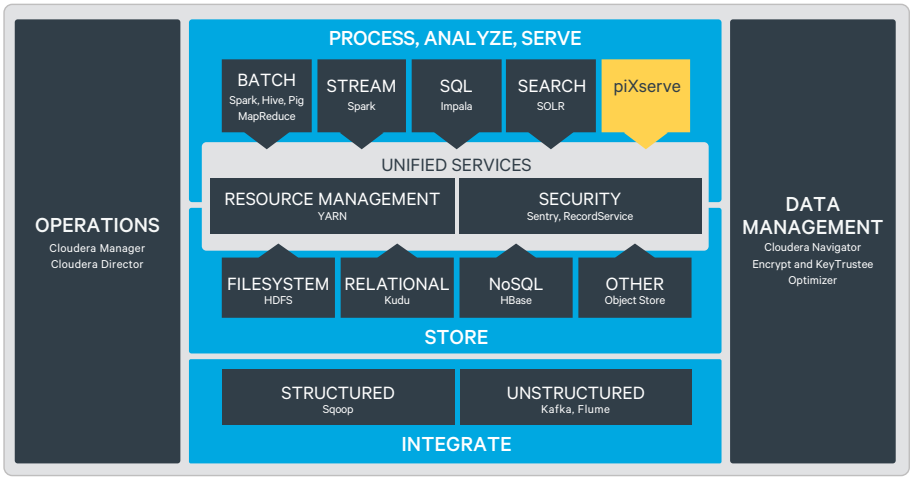
Intel estimates that its processors power more than 97 percent of all servers deployed to support machine learning workloads. They partner with Cloudera to drive open source, enterprise data management, and analytics innovations to continuously drive innovation and growth. With a focus on security, performance, management, and governance, Cloudera Enterprise software is optimized to run on Intel®-based data center technologies to deliver the power and speed necessary for large distributed computing environments.

- **Intel Xeon** – The powerful new Intel® Xeon® processor product family is designed for building next-generation data centers and software-defined infrastructure supercharged for efficiency, performance, and agile services delivery across cloud-native and traditional applications. The processors support workloads for cloud, high-performance computing, networking, and storage and are ideally suited to big data analytics workloads.
- **Intel Xeon Phi Accelerators** – The Intel Xeon Phi processor is a bootable host processor that delivers massive parallelism and vectorization to support the most demanding high-performance computing applications. The integrated and power-efficient architecture delivers significantly more compute per unit of energy consumed versus comparable platforms to give you an improved total cost of ownership. The integration of memory and fabric topples the memory wall and reduces cost to help solve the biggest compute challenges faster.

- **Intel Data Analytics Acceleration Library** – The Intel® Data Analytics Acceleration Library (Intel® DAAL) is used by Cloudera to provide solutions with improved performance and reduce the time it takes to release enhancements. Intel DAAL helps applications make better predictions faster and analyze larger data sets with the available compute resources at hand. The library is updated to take advantage of next-generation processors even before they're available.
- **Intel Math Kernel Library** – Intel® Math Kernel Library (Intel® MKL) is a carefully optimized math library that accelerates math processing and neural network routines. Intel MKL includes highly vectorized and threaded Linear Algebra, Fast Fourier Transforms (FFT), Neural Network, Vector Math, and Statistics functions optimized to run on Intel Architecture to increase application performance and reduce upgrade complexity.

Big Data for Images and Videos

piXserve (the piXlogic product) is a natural fit in the Cloudera solution stack, and with it an important gap in the ecosystem can be closed. Images and videos stored in HDFS can now be on the same footing as other types of structured and unstructured data in the cluster.



Running in the Cloudera stack, piXserve indexes and generates searchable metadata that is used to support content discovery and alerting. End users can quickly search images and videos stored in HDFS, and can do so using a web browser on their desktop or mobile device. Workflows normally associated with images/videos can now run on an HDFS foundation, providing all the data management and security services available from Cloudera. Further, through the piXserve API, the data can also be served to other applications running in the stack. This makes it possible to include media content in the overall scope of analysis, which opens up significant opportunities for identifying correlations across different sources of customer data.

As a full-fledged citizen of the stack, the burden to the cluster administrator of running piXserve is minimal. piXserve worker nodes are deployed on some of the nodes in the cluster. These nodes serve the dual function (fungible) of both indexing media and serving search results. The worker nodes are managed by separate controllers: a piXserve search controller and piXserve indexing controller. In turn, these interact with a piXserve application server outside the stack that interfaces to users.

Indexing pictures and videos in piXserve is a compute-intense operation. Sizing of worker nodes to support the indexing load depends on the volume of media that needs to be processed daily in order to keep up with the flow. Worker nodes need to be powerful and benefit from acceleration (PCIe cards that provide additional compute capacity). The latest-generation Intel Xeon processors, with as many cores as possible, are recommended. When a large pre existing image/video archive exists and needs to be processed, sizing of worker nodes depends on how fast the customer wishes to process that material.

Searching in piXserve is memory bound. piXserve tries to load as much of the metadata it creates as possible in memory. This provides for fast searches. Sizing of worker nodes that support serving search results is based on estimates of the number of concurrent users (users hitting the search key at the same time) and the maximum desired response time defined by the customer.

Since resource allocations are dynamic in a Cloudera cluster, the system administrator can optimize operations to get the most efficient configuration.

Storing all data types in HDFS (text, structured data, images, videos) simplifies workflows and reduces complexity for customers. Being able to carry out more comprehensive and integrated analyses across data types provides new value and opportunities in many applications and use cases.

Using piXserve to help Law Enforcement achieve higher levels of productivity and effectiveness

Images and videos are a mission critical aspect of many law enforcement related workflows and activities.

Use case example 1: Who is this person?

Consider for example, the work carried out by agents at the border. When an undocumented person is caught crossing the border illegally, the agent has little to go by other than the word provided by the individual. Verifying the information provided is a process that can take days and sometimes weeks, as the work is done in large part manually at central locations. The individuals of highest importance are people on watch lists and people that have committed crimes in the past, and these are most likely to provide false information. As a result, it is difficult for field personnel to triage the trespassers that have been caught and discern situations involving carefully planned, multi-point, infiltration schemes looking to get teams of "bad actors" across the border. Other than doing things manually, solutions today have invoked the use of expensive, specialized hardware, that can only check against a limited set of data sources.

Imagine, on the other hand, putting piXserve to work on the back end, processing images/videos from a variety of sources: photos of people on watch lists, repeat offenders, people deported multiple times, visa applications, driver license databases, street cameras, etc. A field agent using a cell phone (no special hardware required), can then at any time take a snapshot of an individual, and retrieve from piXserve in seconds matches that can help the agent understand quickly who he/she is dealing with. Information empowers field personnel. This level of automation can be achieved today broadly and at a low cost.

Use case example 2: How did we get to this?

As another example, consider the situation of videos recorded using body-worn cameras. Today, the video is acquired and stored, and collections are growing. However, for the most part the videos are only used as documentation for the specific situation recorded. These videos could contain information that may be of relevance to other investigations, other situations, perhaps spanning longer periods of time. Today, this information is not mined nor easily made available to the investigative staff. Law enforcement is incurring the costs of acquisition and storage, but not deriving the full benefit of what these assets contain. By processing these videos through piXserve, officers would have the opportunity to discover patterns and correlations, and trace back historically when and where people and objects of interest were observed. The face of an unknown person in the crowd in a video may not mean much in the context of a single criminal event. But if the same unknown face appears similarly in footage from multiple events, it can be an important finding. This is the same for objects, cars, license plates, etc.

Use case example 3: What's going on?

Security cameras are all around us. We feel like we are being watched in parking lots, on the street, in stores, in elevators, at the toll booth, at the airport, everywhere. In reality, however, although cameras are ubiquitous, the monitoring is superficial. There are not enough human eyes to watch that output, and even if we had, even the best human brains can only remember/correlate but a few handfuls of things at a time. The list of people, objects, situations of interest is far too long for a single person to handle. This is where an extra set of machine eyes could really be helpful.

By processing live feeds from security cameras through piXserve, at scale, the videos could be pre-filtered so that what is presented to human eyes for review and decision making would be just the segments most likely to be of interest. For example, if the software identifies several persons of interest loitering in a general area within a narrow span of time, there is a chance that they are all there for a purpose. If a person of interest drops off a passenger at the curb, near a sensitive area, perhaps that previously not noteworthy passenger should also now become a person of interest. These are the type of scenarios that automation can support to make law enforcement personnel more aware and more focused on what could evolve as a significant threat or event.

Conclusion

Hadoop is rapidly becoming a mainstay in organizations due to its flexibility, scalability, and low cost of storing and processing raw data. Cloudera Enterprise—comprising CDH, Cloudera Manager, and expert support—makes Hadoop a reliable, enterprise-ready platform. With an increased focus on improving operational efficiency, leading organizations across industries are moving mission-critical data processing and historical data storage to Cloudera Enterprise. This enables them to store raw data in its native format and develop and maintain complex data pipelines faster, and at significantly lower costs, than were possible using traditional systems. In many customer environments, adding an image/video analytics layer to the workflow can enhance the value of the system even further, creating new opportunities to increase revenue and decrease costs.

About Cloudera

Cloudera delivers the modern platform for machine learning and advanced analytics built on the latest open source technologies. The world's leading organizations trust Cloudera to help solve their most challenging business problems by efficiently capturing, storing, processing and analyzing vast amounts of data. Learn more at cloudera.com.

cloudera.com

1-888-789-1488 or 1-650-362-0488

Cloudera, Inc. 395 Page Mill Road, Palo Alto, CA 94306, USA