databricks

Getting Started With Unstructured Data in Media and Entertainment

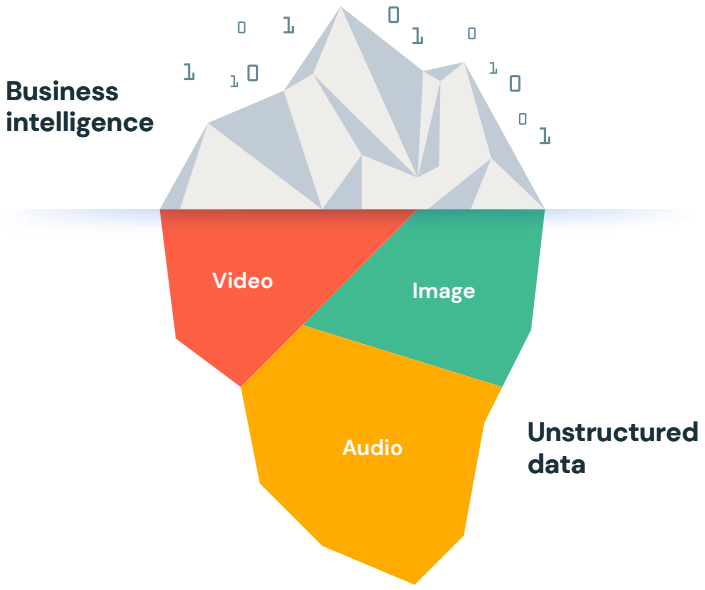
Realize the promise and value of unstructured data in media

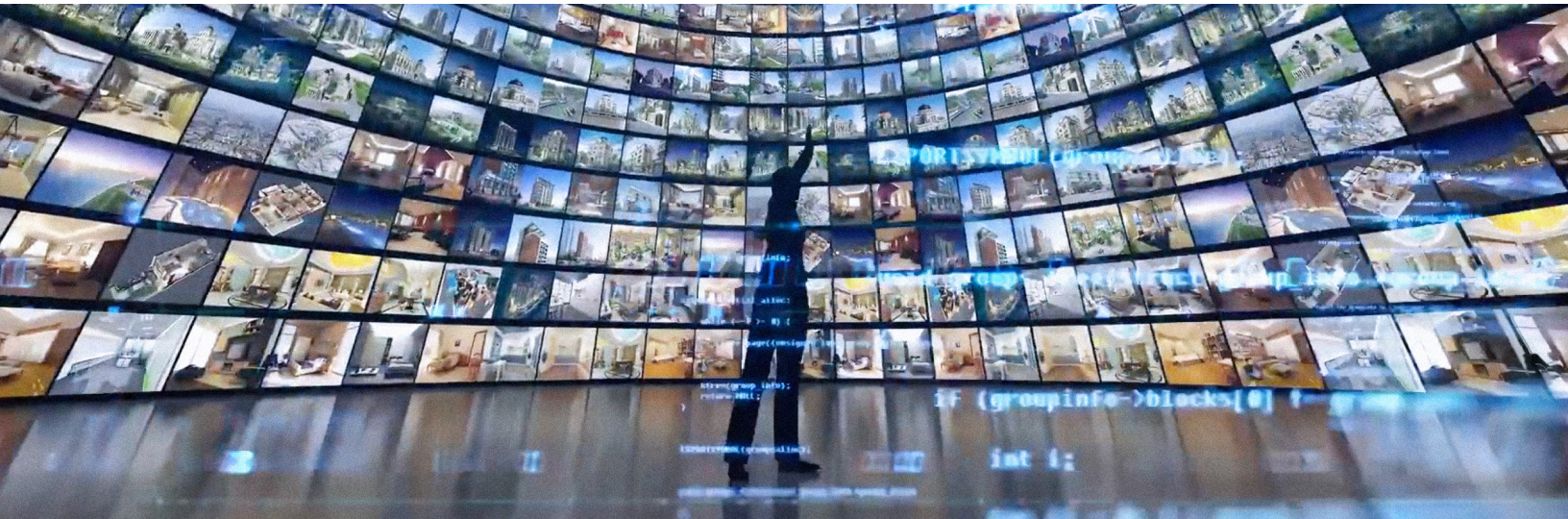
Media and entertainment is the industry of unstructured data (video, images, and audio content). Realizing the full promise and value of unstructured data can make or break companies. While many organizations do an excellent job personalizing the consumer experience and driving greater engagement and monetization of their content, one of the biggest lost opportunities they face is not extracting the full value of their unstructured data.

Media and entertainment companies often have content sitting idle in physical and digital asset management systems. Historically, value is realized at launch, but teams are increasingly finding ways to repackage content, provide recommendations deeper into their catalog, and use machine learning and AI to protect rights and revenues of their unstructured data. They're asking questions such as "Are we leaving money on the table? Are we consistently tracking the rights of my content across channels? Are we empowering our teams to be more productive?" With unstructured data, it's not just about repackaging content that exists, but rather how teams can surround that content with more effective marketing and smarter operations.

In this paper, we dig into common myths about operationalizing unstructured data, as well as dive deep into the most critical scenarios that will make your unstructured data core to your data strategy for years to come.

80% of enterprise data is now unstructured





databricks

Common myths and gotchas in operationalizing unstructured data

Despite the fact that unstructured data forms the foundation upon which the entire media and entertainment industry is built, many organizations still struggle with accessing, analyzing and deriving maximum business value from their video, documents, images and audio files.

Let's take a moment and dispel some of the most common misconceptions related to unstructured data:

My unstructured data is not organized enough to get value from.

Chances are, if you're an enterprise, you've been accumulating data in documents, images and emails for a long time and are likely sitting on a gold mine of untapped potential! What you are probably missing are the tools to explore this value. In the past this has required painstaking work and high budgets for large teams of people to sequentially catalog large data sets using antiquated techniques like "data entry" or expensive techniques like building AI models. This often required high spend on data before even knowing the data's true value. Fortunately with today's technology, it's much simpler. Leveraging Databricks integration with Labelbox Catalog, teams explore areas of value in their data with a fraction of the work by leveraging tools born in the world of machine learning that are now packed for easy use by anyone with an interest in their data.

I'd love to query my unstructured data (i.e., images, documents and videos) using analytics tools like SQL, but creating the metadata requires an ML team we do not have.

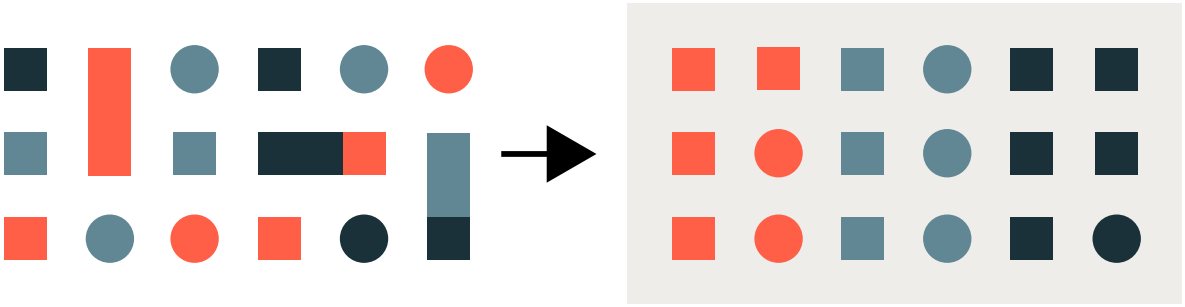
Not all unstructured data requires an ML team to query it! Today, data teams can often leverage tools like Labelbox's Labeling Functions to extract structured metadata without the need for a data scientist or ML engineer to write a single line of code!

I recognize my unstructured data will require an ML dev and data labeling effort, but I'm concerned that the large volume of data will be too expensive to label.

This may feel like a needle-in-a-haystack problem. Fortunately the years of brute force labeling campaigns and giant labeling teams are behind us. Today, data science teams leveraging Databricks and Labelbox benefit from powerful tools like the Labelbox Connector for Databricks and Labelbox Catalog, which help you quickly search through your haystacks of data to pinpoint the needles of business value. Whether you're searching your vast programming content for seasonal holiday visuals, improving search keywords for your end users or placing ad markers to monetize your ad publishing platform, a leading-edge data engine is essential to finding the gold nuggets in your lake of data.

I don't get the sense data platforms offer an easy way for companies like mine to produce labeled data. I need to focus on my business, not my development pipelines.

Fortunately Labelbox and Databricks offer a pre-integrated solution that follows industry best practice. Any team looking to get started quickly, or looking to significantly speed up their labeling efforts, should start here! (No dev skills required.)



databricks

Harnessing the power of unstructured data with Databricks and Labelbox

Customers familiar with Databricks often start by working with structured data. Labelbox is a data engine platform that allows media-rich companies to quickly produce structured data from unstructured data (text, video, document, audio, images). Combining Databricks and Labelbox gives you an end-to-end environment for unstructured data workflows — a query engine built around Delta Lake, fast annotation tools, and a powerful machine learning environment.

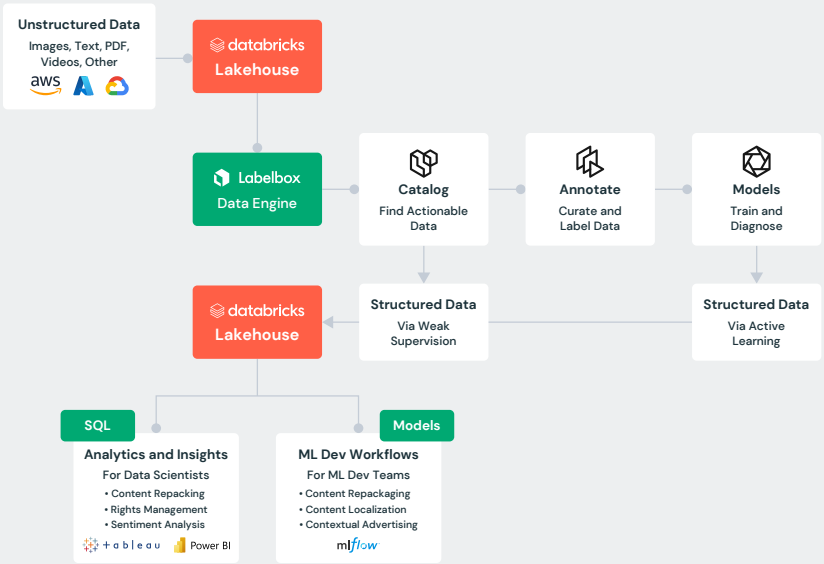
With Databricks and Labelbox:

- Instead of sifting through hours of footage or thousands of images, you can find what you need with a simple query. Once labeled, manual searches that took hours can now be done in seconds.
- You can train your ML to do difficult mission-critical tasks with unstructured data, such as find and remediate toxic behavior on social platforms to enhance brand safety or improve diversity in content acquisition and production
- Automation through model-assisted labeling has allowed companies to reduce data labeling time by 50%-80% while improving ML model performance on less training data. Teams can produce more accurate models in less time with Databricks and Labelbox.
- Data analysts and BI teams can now self-serve when curating their unstructured data, providing quick value without having to wait for an AI/ML team to build an expensive model
- Do what your competition cannot: Train ML models on your proprietary, unstructured data at massive scale. Producing your own data sets gives you a competitive advantage over your peers. And with practically unlimited compute in Databricks to query and derive insights from your data sets, there are no limits to what your BI and AI teams can unlock!

Databricks and Labelbox architecture

Combining Databricks and Labelbox gives data and AI teams an end-to-end environment for unstructured data workflows, along with a query engine built around Delta Lake, coupling fast annotation tools with a powerful machine learning compute environment.

Lifecycle of data | Workflow overview



HOW IT WORKS

The Labelbox Connector for Databricks is a Python library. Use it to register unstructured data from Databricks in Labelbox to label/annotate your data.

- Easily pass annotations back into Databricks so you now have your training data in your lakehouse. Feed it to your models, cloud AI interfaces or both.
- Visualize model errors and take action to improve performance faster with Model Diagnostics in Labelbox (Model Diagnostics works seamlessly with MLflow)
- Prioritize the data that will have the biggest impact on model performance with Catalog in Labelbox

databricks

Content repackaging brings speed and agility to content creation

Every company with a vast library of content is continually seeking new ways to extract value from the content they already have access to. For example, in the moment that a football player announces retirement, an opportunity for engagement awaits any outlet that quickly publishes a sizzle reel highlighting the player's career achievements.

In the past, creating and publishing this type of asset required having people manually sift through a large amount of content, guided by memory or high-level content descriptions. Today, with media assets readily accessible in the cloud, leading companies are creating these types of assets just in time to effectively monetize social buzz.

PRO TIPS

- To repackage content effectively, ensure that you have in place a strong communication and collaboration workflow to get your data engineers and labeling workforces all working in tandem to ensure quality.
- A clean, thoughtful ontology is critical for creating high-quality labeled data with minimal errors for your content. Create an ontology that follows the most logical workflow for your data science and labeling teams.
- Once your initial model is trained, you can leverage comparative analysis between your ground truth and predictions from your AI models to find and fix human labeling errors so that your content is properly labeled.

databricks

Contextual advertising unlocks new revenue opportunities

For decades, companies have relied on behavioral ad targeting to reach their target audience. With this technique, an automotive company, for example, could direct their online advertising to people who have demonstrated intent to buy – deduced by their web browsing activity. While this method of advertising has been effective historically, its future remains uncertain as the privacy landscape rapidly evolves (e.g., GDPR, CCPA, Apple's App Tracking Transparency Framework, deprecation of third-party cookies).

Enter contextual advertising.

In contrast to behavioral ad targeting, contextual advertising delivers ads based on the context in which they are viewed, as opposed to the browsing history of a given user. For example, the previously mentioned automotive company could reach their target audience by delivering ads against reviews of cars in the same class as their own or against premium video content that includes similar vehicles. As demand shifts to this form of advertising, media companies are now annotating their content libraries (e.g., video, images, text, audio) to unlock new revenue opportunities. Facilitating this at scale is the IAB's Content Taxonomy, which standardizes the labels that are used when transacting within the media ecosystem.

PRO TIPS

- Model embeddings help you quickly uncover high-level patterns and visually similar data from across all your data sets. This process allows teams to search across millions and millions of data points in minutes rather than weeks.
- Leverage workflows such as weak supervision to automatically apply labels, metadata and insights across your unstructured data without needing to always build a model.
- When it comes to ad content, be sure to utilize all your metadata and text embeddings so that you can automatically add labels at scale and queue them for human review.

databricks

Model-assisted labeling enables content localization across regions

Global presence is a critical part of every media company's growth strategy. At the same time, producing content that is exclusive to each and every region isn't a tactic that scales well given the cost of content production and the difficulty in producing that next new hit.

Enter content localization.

With content localization, media companies take content that was produced for one region and edit it for use in another region. To do this at scale, leading media companies develop workflows for transcribing audio transcripts from one language to the next and then carefully pair the new audio track to the existing video frames. Further, care is taken to make sure that cultural references carry over appropriately from one region to the next.

For example, a Fortune 250 Media and Entertainment company is creating closed captions for various business videos that are being sent across the globe. These closed captions are mapped to specific videos and then corrected for each local geography. Leveraging model-assisted labeling to detect such examples is a powerful way to identify such instances and in turn help accelerate global expansion.

PRO TIPS

- Labeling vast quantities of image and video data quickly, efficiently and accurately is an immense challenge, but ML teams have found a way to cut both labeling time and costs with an innovative solution via labeling automation. Labeling automation leverages open source models and internal models to increase labeling efficiency by more than 50%.
- For labeling tasks that are expensive and time consuming, use your own model to pre-label the data. Once the pre-labeling is accurate enough, the task can be turned over to less skilled and less expensive resources to verify the automated labels, greatly reducing time and expenses. This kicks off the active learning lifecycle.
- Use publicly available open source algorithms to automate the labeling around common objects. No longer do teams need to spend resources and time labeling cars, people and plants when model-assisted labeling can help you achieve this outcome faster.

databricks

Computer vision simplifies rights management and royalty reporting

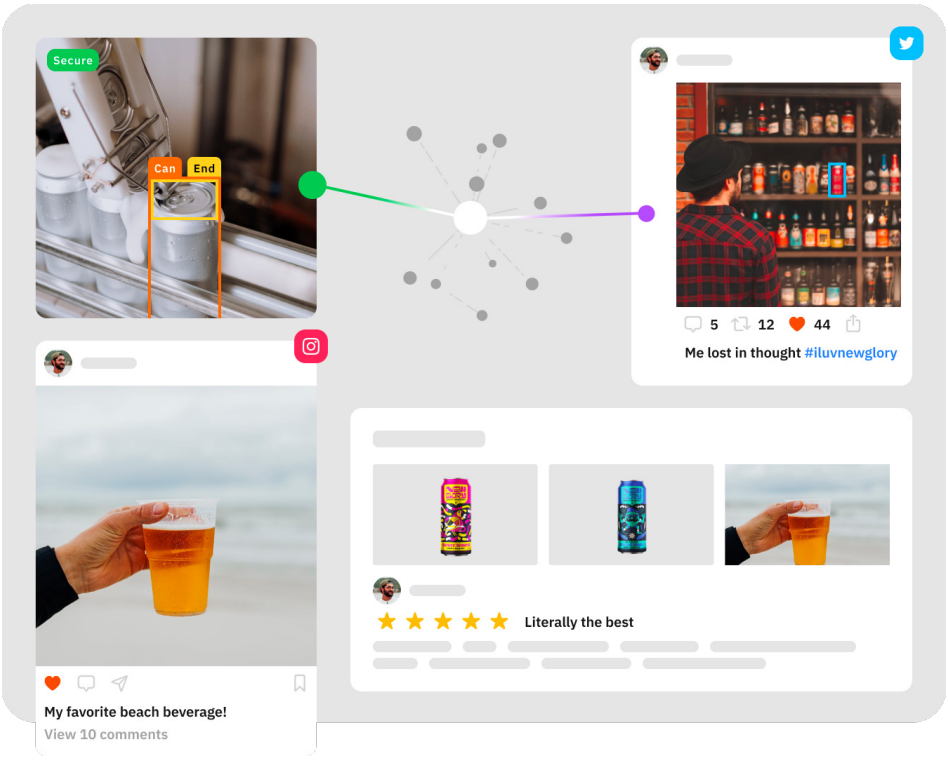
Building an instantly recognizable brand that is loved by millions is incredibly hard. For the companies that succeed in doing so, however, opportunities to further monetize brand affinity await in the form of licensing. Such licensing agreements can result in new content, experiences and consumer products. The details for these agreements can get quite complex.

For example, in the case of animated characters (e.g., Spider-Man, SpongeBob), royalties are often accrued for the amount of time that the character appears on screen. But what if only the hand appears? Or maybe just a small fraction of the character? Such examples are in fact real and need to be taken into account when calculating royalty payment. Computer vision is the key to doing this efficiently at scale.

PRO TIPS

- To get started faster, set up a workflow to quickly search and visualize all of your unstructured data in one place. With all your data, metadata, labels and predictions at your fingertips, you can make better decisions for streamlining rights management.
- For licensing agreements, focus on how you can leverage your existing repository of PDF documents, which are inherently complex. They often contain lots of text, images, charts, graphs and more. ML teams should annotate text of interest alongside OCR techniques to effectively capture both content and context.
- Find ways to prevent duplicate data, which can be crucial as your projects scale to millions of documents. This is accomplished by setting up processes to easily handle data row imports, querying, and deleting data rows in a single interface.

Sentiment analysis enables informed decision-making



Consumers and influencers have more ways than ever before to share with the world their sentiment for virtually anything, including the content, products and experiences that are released by media and entertainment companies. This type of information is immensely valuable for organizations that effectively harness it.

For example, the massively multiplayer online games (MMOs) released by game studios are incredibly complex and at any given time, there are inevitably bugs to address or features to improve. By applying natural language processing to the comments posted in game forums, many studios are able to quickly identify issues to address. This information is then incorporated into the game studios' product road maps.

Outside of gaming, PR agencies use this type of data as well to establish a PR rating, or measure of trustworthiness, for popular influencers. These measures are then used to pair advertisers with influencers as a means of cultivating buzz for a given product.

PRO TIPS

- To accelerate your progress, utilize open source models and MLflow to kick off sentiment analysis inferences across your database. Use Active Learning techniques to utilize human intervention only on the lowest confidence data rows.
- Data that a model is uncertain of are also often more engaging to annotate and more important to review, allowing the human layer of your machine learning system to spend their time on high-impact decisions.
- Set up a way to easily search for text data using filters such as annotation, metadata and similarity embeddings to prioritize text snippets to label or create review tasks to fix issues that matter the most.

databricks

Customer references

Discover how innovative media and entertainment companies are leveraging Databricks and Labelbox to unlock the power of their unstructured data.

DoubleVerify

DoubleVerify's Semantic Science team builds AI-driven ontological tools that facilitate the semantic understanding of text. This unique competency powers DoubleVerify's proprietary brand safety and suitability controls for advertisers and matches brands with appropriate and relevant content online.

CRITEO

Criteo, an online advertising enterprise powered by AI, struggled without a reliable data labeling pipeline and collaboration method. They often relied on emails and spreadsheets to communicate and track their labeling, which quickly became untenable. Once they invested in Labelbox, the company realized "humongous benefits," according to Hong Noh, Sr. Product Manager at Criteo. These benefits include a 40% increase in annotation speed, improved label quality, and a massive reduction in back-and-forth emails.

SPORTS LEAGUE

A major international sports league was struggling with the scale and complexity of in-game player analysis use cases that were coming from in-stadium cameras. They are now able to deliver real-time use cases for pre- and post-game player analysis looking at weather conditions and player matchups as well as predicting player injuries before they occur.

KEY TAKEAWAYS

- Media teams have vast amounts of unique and proprietary unstructured data. Once these use cases have structure, it unlocks the value for consumers, internal teams and overall the opportunity to drive more revenue across channels and partners.
- While unstructured workloads were primarily the domain of data science teams, with complexity of tooling being a primary barrier to entry, data analysts and BI teams now have access to platforms that enable these key personas to derive benefit from handling unstructured workloads.
- The surface area of use cases for unstructured data is only limited by the use cases you want to solve. As 80% of all data in enterprise is now unstructured, there is a significant opportunity for organizations to drive more value and more use cases aligned to these critical workloads.
- Extracting insights from unstructured data doesn't need to be hard. Databricks and Labelbox offer a fully integrated solution that brings together a best-in-class labeling platform running on top of the default standard for enterprise data processing and ML.

Interested in getting started for driving more value around your video, images, text and audio files? Check out the How to Get Started section, which links to technical assets that you can use to deploy production workloads in days and weeks, not weeks and months.

databricks

How to start unlocking the value of your unstructured data

The flexibility of using Databricks and Labelbox together means that you can start with the use case that will be most impactful for your business. As you implement the pattern, you will find that you're able to tackle use cases quicker and more easily than before as you more rapidly drive value for the business.

Connect to Labelbox

Databricks and Labelbox have collaborated on a library of technical guides that will allow you to kick off active learning, weak supervision and model diagnostic workflows in minutes. With these step-by-step instructions, we show you how to configure your environment and get started with unlocking the value of your unstructured data today.

Get Started Today

Related resources

Productionizing unstructured data for AI and analytics

Databricks integrates data tools with Partner Connect

Burberry keeps standing out with Databricks and Labelbox

Learn more

Discover more about the Databricks Lakehouse for Media & Entertainment