Accenture Applied Intelligence

UNLOCK VALUE FROM BIOPHARMA ENTERPRISE DATA LAKES

2018

Using search and analytics to accelerate breakthrough innovation, efficiency, and speed to market

In recent years, the demand for faster, more efficient data access and analytics at end-users' fingertips has fueled the emergence of **enterprise data lakes (EDLs)**: repositories designed to hold vast amounts of raw data - unstructured and structured - in native formats until needed by the business. But implementing an EDL is just a start. How can organizations unlock its full value?

80% of enterprise data is unstructured (email, videos, lab notes, contracts, research, etc.), but most analysis has been done only around structured data (spreadsheets, financial records, product names and numbers, customer names and numbers, etc.). To unlock the full value of the EDL, the key is the ability to derive a 360-degree view of the enterprise from both types of data.

A RECENT STUDY

Aberdeen Group found that volumes of data in the average company are growing by over 50% per year. Additionally, these companies are managing an average of 33 unique data sources for analysis.[1]

Like every business in the era of digital transformation, biopharmaceutical organizations rely on diverse data-driven applications and content repositories to support their day-to-day operations and strategic goals. Every function along the pharmaceutical value chain – from R&D, clinical trials and manufacturing, to distribution, pharmacovigilance, and regulatory compliance – must deal with massive amounts of data from disparate unstructured and structured sources. The bottom line? Without the right technology and approach for managing and deriving value from data, it will be challenging – sometimes impossible – to solve difficult problems and gain a competitive advantage. Finding answers to the biggest questions starts with data.

EDL users:

- Scientists
- Researchers
- Business Analysts
- Technical Architects
- Project Managers
- Program Analysts
- Information Systems Managers
- Information Systems Architects

This e-book explores search and analytics applications built on top of EDLs and how they can help biopharmaceutical companies accelerate innovation and business success.

Key areas covered include:

- the evolution of the data lake
- benefits
- search and analytics applications
- reference architecture
- deployment process
- measures of success

DID YOU KNOW?

In an Aberdeen survey, organizations that implemented a data lake outperformed their peers by 9% in organic revenue growth.[1]

The evolution of the data lake – toward new performance and scalability

Enterprises in every industry have been playing the data game for more than three decades. The priorities? Bringing together multiple sources of data assets into a usable format, before serving it up to business users at the right time, in the right place, and in the right format. Data lakes are just the next phase in the evolution of data repositories being used for various business purposes.

Relational databases

Starting in the 1970s and becoming widespread in the 1980s, relational databases based on Structured Query Language (SQL) were the primary vehicles for enterprise data storage and analysis. In life sciences companies, they were used mainly in finance and business operations and later on to store manufacturing data for compliance purposes. But these databases had some major drawbacks. Along with high costs, they carried rigid structural requirements that resulted in data type limitations and even data losses. At the same time, a proliferation in departmentalized databases to support individual enterprise functions marked the onset of siloed databases that persist in many organizations to this day.

Data warehouses

During the 1990s, fueled by the advance of the internet, greater access to computers, and increasing digitalization, enterprises across all industries experienced exponential data growth. With more data sources, expanding data volumes, and increased pressure to discover business intelligence from that data, data warehouses emerged as the central enterprise data repository. Although they were useful for aggregating data for business analytics purposes (e.g. plant yield), they were less useful for operations. And along with high price tags, they required specialized knowledge to access and analyze data. What's more? Because all data had to be standardized before going into the data warehouse, information discovery and analysis was often a daunting and lengthy process.

Enterprise data lakes (EDLs)

The rise of cloud in the past decade has provided cheaper, more flexible storage for the explosion in big data that's taken place. The creation of big data repositories based on Hadoop Distributed File Systems (HDFS) thrived as organizations sought to infinitely scale their data applications. As an advanced alternative that overcame the limitations of traditional relational databases and data warehouses, the EDL answered calls for transformative ways to gain faster, better business intelligence from ever-growing data – whether unstructured, structured or semi-structured. For the first time, operational and business data could be stored along with new data sources and made accessible via different applications and interfaces that break down silos.

The EDL centrally stores data from hundreds of sources in their raw format, deferring analysis until it's needed by the business. When analysis begins, the EDL will only pull in, analyze, parse, and clean up the data that's required for searching, reporting, visualizing, and troubleshooting. The analysis can produce additional processed data that is fed back into the EDL. This creates a feedback loop: raw data goes into the lake for analysis and processing, and post-processed data is put back into the lake for potential further analysis.

In the life sciences space, specifically for biopharmaceutical companies, R&D, collaboration, manufacturing, and regulatory compliance are among the many data-intensive areas that EDLs can support efficiently and cost-effectively.

Enterprise Data Lake vs. Data Warehouse

Unlike data warehouses, EDLs have a number of unique capabilities. They can:

- ingest all raw data
- retain all data indefinitely
- analyze data only as needed
- scale indefinitely

Enterprise data lakes: the benefits

For the foreseeable future, legacy data environments will continue to perform targeted functions for life sciences companies. As well as high-volume ad-hoc querying, there will be financing, risk, and reporting applications. For other use cases, however, it will make sense to quickly migrate from old platforms, minimizing the cost and user overhead involved and reducing the potential for operational risk. As organizations migrate to EDLs, they will experience a number of major benefits:

Centralized content silos

- Enabling better data discovery through search and retrieval of content in the EDL. Often, source systems are deployed in separate geographical locations, where bandwidth, access, or other constraints can prevent direct access. However, the data lake copy of the content makes it accessible to EDL users regardless of the source system's location.
- Extending content access beyond the constraints of the source system:
  - There may be a limited number of seats or licenses for access to the source system.
  - Some users may only need access to the content in a given source system for short-term activities, such as research and due-diligence.
  - Users from the same department or unit should have access to all content from that part of the organization regardless of the source system.
- Providing better mechanisms for collaborating with other EDL users across multiple content sources.

Overcome legacy system limitations

- Avoiding risks of disruption when the source system is out of support by the manufacturer:
  - The platform supporting the source system may be too old
  - The servers hosting the source system may be offline (after the content was copied to the EDL) and only restored or turned back on temporarily under specific circumstances

- Data enriched in ways not possible in source systems
- Cleansing and normalizing metadata
- Enabling identification of similar data across different sources, despite some differences in the data between one source and the others, or even within the same source
- Tagging and classifying semi-structured or unstructured data

Analytics transformed

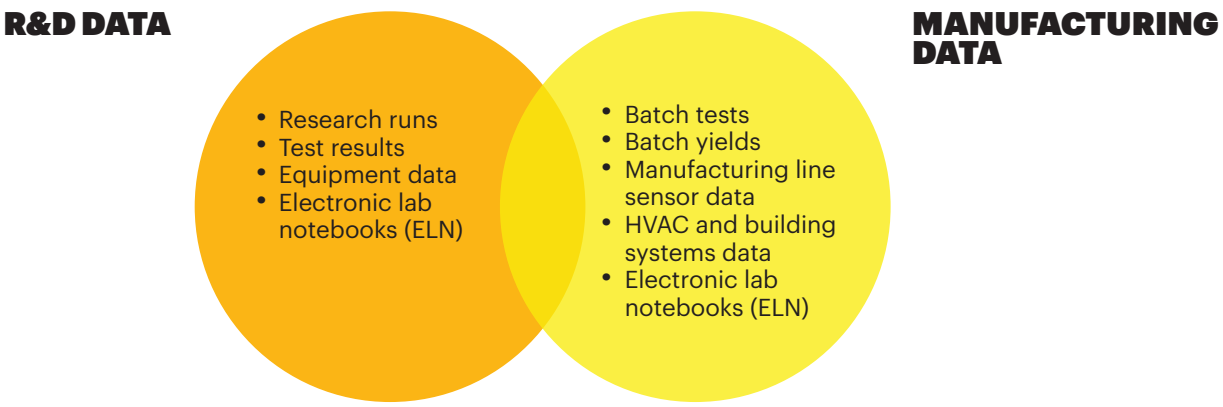
- Facilitating traditional reporting, dashboards, or similar analytics functionality
- Enabling more modern analytics capabilities implementation, such as forecasting, prediction, and machine learning
- Unlocking insights from both unstructured and structured data

Maximizing value from data lakes with search and analytics applications

The end-goal of EDLs is to provide a centralized repository of structured and unstructured data so that business users can access, search, and conduct analytics in real-time.

For instance, there are tremendous benefits in being able to have a single, 360° view across all life sciences content sources. Such sources could include Documentum, IBM Connections and SharePoint, along with Electronic Lab Notebooks (ELNs) like IDBS E-WorkBook ELN, CambridgeSoft ELN, NuGenesis ELN by Waters, network file shares, and third-party data (e.g. data from ClinicalTrials.gov or other relevant industry databases). Since these are separate repositories with different formats and configurations, bringing together data from these sources can help increase visibility into Manufacturing and R&D processes – the two critical operational areas for biopharmaceuticals.

Below are some examples of the data types needed to support R&D and Manufacturing functions, including analyzing root causes, reducing drug discovery time, eliminating research duplication, and avoiding non-compliance risks.



Examples of EDL search and analytics applications

Life sciences applications that can be built on top of an EDL include:

Cross-silo search: an enterprise-wide search application that supports both R&D and Manufacturing operations by enabling users from different enterprise departments and geographical locations to have flexible access to the data lake's content from anywhere. This helps encourage knowledge sharing, avoid duplication of effort, and provide an efficient way to gather the right data required to drive business decisions.

Legal search: an application that enables lawyers to search across legal contracts and compliance documentation based on the metadata associated with the content. This is critical to ensure compliance and avoid legal risks.

Video search: an application for searching across text transcripts from training videos (e.g. clinical surgeries, manufacturing procedures) generated by employees and contractors.

Molecule search: a feature in both Manufacturing and R&D search applications that takes a molecule and searches for experiments that exactly match the query molecule, contain the query molecule as a substructure, or contain a molecule similar to the query molecule. Molecule searching may be done using, for example, SMILES notation (Simplified Molecular-Input Line-Entry System) for the representation of the indexed molecule. The search user interface may also allow the user to draw a molecule, which would be converted as the query molecule in the same representation used in the index, such as SMILES. Searching by molecules can lead to the discovery of other experimental data done with the same molecule, with that molecule as a substructure, or with similar molecules. Experiments found via molecule search can help avoid duplicate experimentation, provide information that may change the design of an experiment or accelerate new experimentation, and identify additional use cases for that molecule.

Lab instrument data search: an extension of the R&D search application that supports additional content from network file shares (human- or machine-generated) and focuses on metadata for images or proprietary formats.

Medical insights: an index of interactions, surveys, and discussions among medical practitioners about the usage and effects of life science companies' products in the market. The index aggregates content from various separate content sources and enriches it with identified entities found within the unstructured text or metadata (such as diseases and products).

Reference architecture

Core components

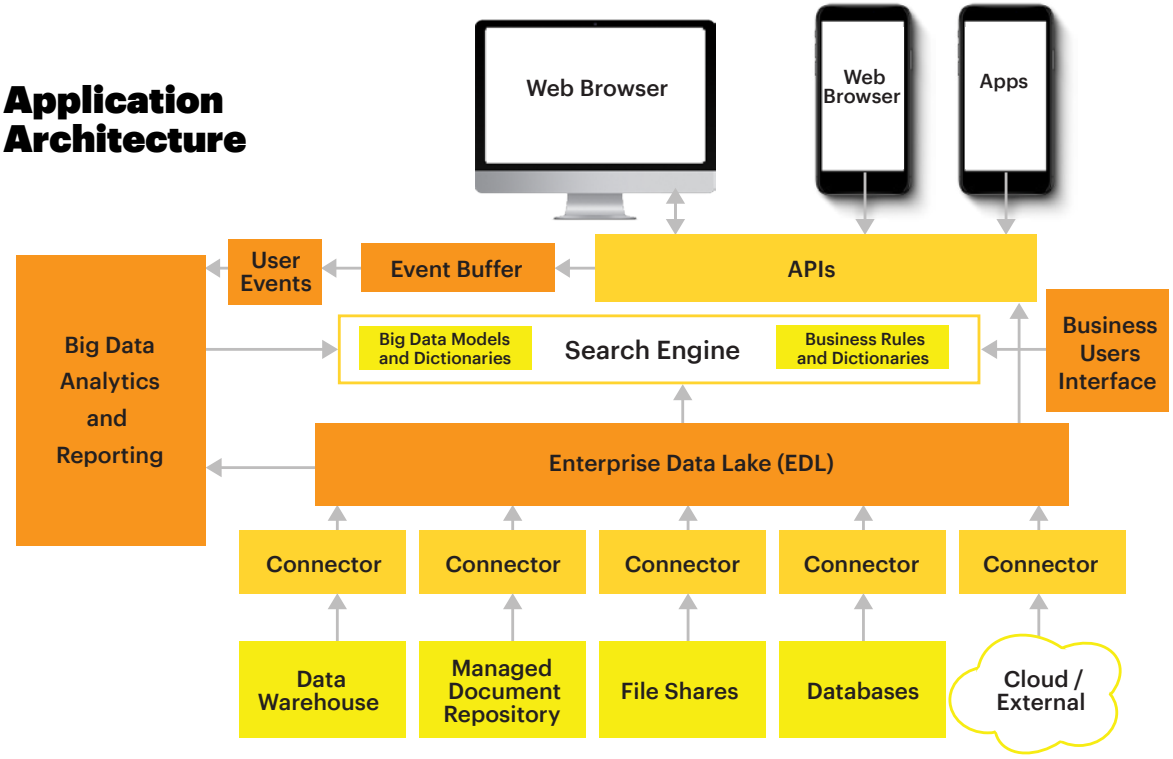
From a conceptual architecture point of view, the fundamental feature of an EDL is sharing. Data must be sourced from multiple sites and brought into the data lake. Abstracted layers of data then have to be instantiated to provide data

to consumers in suitable formats for uses including reporting, ad-hoc querying, and advanced analytics.

The graphic below shows a reference architecture for EDL applications. Its core components are:

1. Connectors to original content sources
2. Data storage repository
3. Search engine
4. APIs to connect the applications to different enterprise applications and user interfaces
5. Big data analytics and reporting algorithms

Application Architecture



All content from **disparate repositories** is ingested into the EDL using **connectors,** stored in its original format, and made available to users via a search engine. **Data enrichment algorithms,** based on big data models and dictionaries, are leveraged to enrich the data prior to indexing it to a search engine.

A search engine is the ideal tool for managing the data lake because:

- Search engines are user-friendly.
- Search engines are schema-free – schemas do not need to be pre-defined. Search engines can handle records with varying schemas in the same index.

- Search engines naturally scale to billions of records and perform real-time analytics across that massive amount of data at a reasonable cost.
- Search can sift through wholly unstructured content.

Where necessary, content will be analyzed and results will be fed back to users via search across various **user interfaces,** such as web browsers and mobile apps, within the enterprise.

Commonly Used EDL Technologies

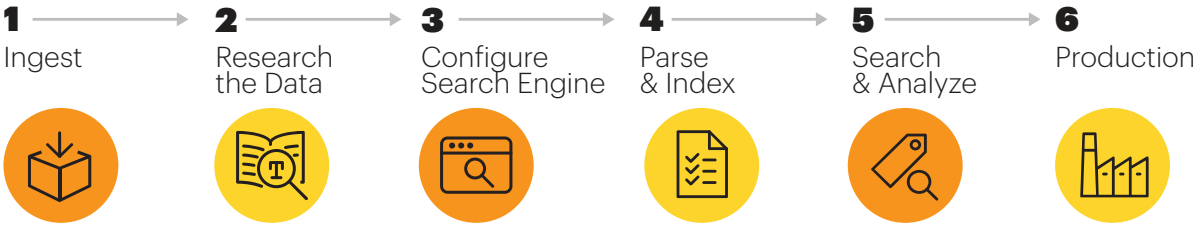
An EDL application can be built on a wide range of technology platforms, both open source and commercial. Below are some technologies commonly leveraged in EDL implementations. Depending on an organization's system and business requirements, the configurations can be customized for peak performance:

- **Search engines:** Solr, Elasticsearch, Microsoft Azure Search, Cloudera Search
- **Content processing and connectors:** Apache Tika, Tesseract OCR, ImageMagick, Aspire Content Processing and Connector Framework, Natural Language Understanding (NLU) or Processing (NLP) toolkits or libraries, etc.
- **Big data storage and processing:** HDFS, HBase, Impala, Morphlines, YARN, Spark, MapReduce, etc.
- **Cloud-based tools and environments:** IBM Cloud, Microsoft Azure products, Amazon Web Services (AWS) products
- **Other:** Java Database Connectivity (JDBC), Java CIFS (JCIFS)

Deployment process

How can EDLs help users find the data they need across tens of thousands of databases and billions of records?

To do this systematically and efficiently, a well-defined process for deployment and management must be in place to ensure that data is available and properly tagged for good search experiences. The following chart shows a common data lake process flow in which specific tasks can be created to fulfill unique organizational requirements.



1. **Ingest** – pull data in raw formats from the original sources into the data lake, storing in HDFS, HBase, Impala, etc.
2. **Research** – understand the format (.csv, .doc, .pdf, etc.) and semantics of that data
3. **Configure** – with an understanding of the data, the search engine fields can be configured and processed correctly to support advanced search features
4. **Parse, index, and clean-up** – to make data available and reliable for search and analytics
5. **Search and analyze** – the indexed data can be searched and displayed as text results, real-time charts and graphs, or a combination of both based on enterprise requirements
6. **Move to production** – once in production, multiple tasks should be repeated on a regular basis to ensure process efficiency and data quality. These tasks could include testing and quality assurance (QA), incremental processing, scheduled workflow, and security controls.

Measures of success

As data-driven businesses, life sciences organizations measure the effectiveness of their EDLs based on certain Manufacturing and R&D benchmarks specific to both the industry and the individual organization.

So what does a successful life sciences EDL look like? Here are some examples:

- Supports the **analysis of historical product performance** and quality information across different products with similar profiles

REDUCE TIME, IMPROVE INSIGHT DISCOVERY

Root cause analysis:
Weeks → Minutes

Query time:
Days → Minutes

Time to make external data available for analytics:
Weeks → Hours

- Predicts the **likelihood of successfully manufacturing additional batches** without extending production schedule
- **Investigates potential contamination** in the environment and narrows down specific time windows and batches being manufactured in the facility during that time
- **Detects anomalies and identifies root causes** of manufacturing defects

EDLs also bring tremendous value in helping companies solve unforeseen problems, for instance:

Weathering a hurricane

What happened when a devastating hurricane caused critical damage to the information system infrastructure at a biopharmaceutical manufacturing plant? The EDL provided a critical backup capability and allowed manufacturing processes to continue. The EDL search application enabled users to find and download hundreds of experiments needed to certify that the manufactured product met stringent industry standards. Without this certification, the manufactured product would have had no commercial value. In this scenario, the EDL directly impacted the company's production capability and revenue.

Pinpointing external causes of defects

By searching the EDL's content, including unstructured lab notes, a biopharmaceutical company identified certain environmental factors affecting its product quality. With multiple manufacturing facilities, even a minor difference in a facility's lighting or temperature could result in defective products. The EDL enabled users in dispersed manufacturing locations to locate and eliminate manufacturing anomalies by ensuring all environmental variables were uniform across the board.

Are you unlocking the full value of the data lake?

Given the amount of data that life sciences organizations work with, the ability to capitalize on all data, including unstructured and third-party data, is crucial to operational efficiency and business outcomes. EDLs play an essential role in helping organizations gain a real-time, 360° view of their businesses, ultimately improving insight discovery and fueling competitive advantage. Finding the right skills and implementation partners is key to a successful data lake strategy.

Learn more about our Data Lake Solutions and Services, including:

- Strategy consulting
- Data acquisition and preparation
- Architecture design and implementation
- Security and governance
- User interface development
- Managed services

Contact us to discuss how you can leverage an EDL for various life sciences applications within your organization.

DID YOU KNOW?

According to IDC, even though 73% of companies intend to increase spending on analytics and making data discovery a more significant part of their architecture, 60% believe they lack the skills to make the best use of their data.[2]

Watch our on-demand webinar for a deep-dive into searching the data lake to provide a 360⁰ view of the business.

References

[1] Michael Lock, "Angling for Insight in Today's Data Lake"

[2] Prashant Tyagi and Haluk Demirkan, "Data Lakes: The biggest big data challenges"
Amber Lee Dennis, "Data Lake 101: An Overview"
Keith D. Foote, "A Brief History of the Data Warehouse"

ABOUT US

Search Technologies, now part of Accenture, is the leading technology services firm specializing in the design, implementation, and management of search and big data analytics solutions. We bring a deep understanding of the nature of structured and unstructured content to extract knowledge and untrap business value using search and big data. Visit **www.searchtechnologies.com** or contact us at **info.stc@accenture.com** to learn more.

ABOUT ACCENTURE

Accenture is a leading global professional services company, providing a broad range of services and solutions in strategy, consulting, digital, technology and operations. Combining unmatched experience and specialized skills across more than 40 industries and all business functions – underpinned by the world's largest delivery network – Accenture works at the intersection of business and technology to help clients improve their performance and create sustainable value for their stakeholders. With more than 442,000 people serving clients in more than 120 countries, Accenture drives innovation to improve the way the world works and lives. Visit us at **www.accenture.com**.