

# Big Data Overview

---

## The Pharmaceutical Industry



# Pharmaceutical Industry Overview

---

## Status of Big Data Initiatives

The Pharmaceutical industry is becoming a leader in Big Data initiatives as more data sources and use cases are becoming available. With the creation of the HITECH Act, hospitals are being forced to switch over to electronic medical records (EMR). This is providing pharmaceutical companies with an influx of potential data sources that they did not have previously. Therefore, the pharmaceutical industry is now in a position to be able to process and analyze large amounts of patient data if they put Big Data strategies in place. Previously, many pharmaceutical companies were already using other types of Big Data to make faster and more informed decisions within individual business units.

*“The pharmaceutical industry is now in a position to be able to process and analyze large amounts of patient data.”*

Over 90% of Fortune 500 pharmaceutical companies already have Big Data strategies in place and are using some form of Big Data gather platform. Due to the number of divisions within each pharmaceutical company, there is usually a separate Big Data strategy within each department. Therefore, in an organization with twelve business units, there are at least twelve Big Data strategies. Individual business units have ownership over their own Big Data strategies, data, platforms, and tools. Over 50% of companies are utilizing Hadoop within at least one division of the company. Strategies in this industry vary widely by company and by department but the most common Big Data usage is for Research and Development.

## Types of Data

The pharmaceutical industry shares the overall goal of improving patient outcomes through the usage of Big Data. To do so, pharmaceutical companies have embraced

multiple forms of patient data often referred to as Real-World Datasets. The most common form of patient data is from medical records. However, it can also come from unconventional sources such as social media. 50% of data that pharmaceutical companies use is external and companies are interested in accessing and utilizing even more. Pharmaceutical companies have a wide variety of data sources that they can use to piece together patient profiles to improve drugs and processes. This can lead to improving overall health outcomes. While pharmaceutical companies also have a large amount of internal data such as financial, operational, and employee-related, the focus of the industry has recently shifted to new opportunities with Real-World Datasets.

### **Electronic Medical Records**

The most common type of data that 63% of pharmaceutical companies are utilizing is electronic medical records. As electronic medical records become standard across the healthcare industry, researchers will be able to access both structured and unstructured data on patient's medical histories. This will provide insights into patients and diseases. Previously, pharmaceutical companies could not gain that level of insight into patients because they could not readily access their medical histories.

Over 50% of the data that pharmaceutical companies use is public data. The reason why these Real-World Datasets are so important is because companies are not limited to their own information. They can see how diseases are affecting people's everyday lives and are not restricted to the potential biases of a lab setting. Researchers can access a patient's prescription history and see how they are reacting to certain drugs. They can access hospital admission and discharge information to see how long a patient was in the hospital and why. These Real-World Datasets can make a huge impact on the research and development of new drugs and therapies.

### **Clinical Documents and Laboratory Journals**

Clinical documents are unstructured notes written during clinical testing. These documents hold insight into drug interactions in a clinical setting which can help improve clinical trial design and protocol. Similarly, as more researchers are using electronic

laboratory journals there becomes an opportunity to use Big Data analytics to generate insights from laboratory data quicker. The notes can be combined and processed using Big Data analytics to draw insights quicker and more accurately. Pharmaceutical companies can combine laboratory and clinical data from internal and external sources to discover new drugs and therapies.

### **Emerging Data Sources**

One emerging data source for the pharmaceutical industry is social media data. 18% of pharmaceutical companies are already working with this type of data. Social media data can be used to better understand the patient, their medical worries, symptoms, medicine side-effects, medical needs, and drug compliance. These insights can help companies build detailed patient profiles and learn how to improve drugs and therapies to better meet each patient's needs.

Genomic data is becoming an increasingly important data source. Sequencing a single person's genome now costs around \$3,000 and generates about 3GB of data, making it increasingly realistic for pharmaceutical companies to access this data. This genetic data can help predict potential illnesses, target treatments, and monitor medication effects on patients with specific genetic makeups. By combining genomic data with patient medical records, researchers will be able to derive insights into individual patient's genetic makeups and develop drugs and therapies better targeted to their needs.

A third emerging data source is imaging data. Imaging data can come from sources such as MRIs, CT scans, and X-rays. By 2015, 80% of hospital patient data will be unstructured image data. As the amount of patient data increases so do medical imaging archives at a rate of 20-40% per year. Imaging data is unstructured, complex, and difficult to process but when combine with other patient data it can provide important insights into diseases.

## **Data Storage**

A single patient generates approximately 1GB of data in a hospital setting each year. Therefore, between the healthcare and pharmaceutical industries, there is hundreds of exabytes worth of data that is being stored. Most of this information is highly sensitive and the subject of privacy and legal concerns. 63% of pharmaceutical companies rely solely on data warehouses due to privacy concerns. Also, many companies are using separate warehouses for separate data. Much of this data is not being accessed because of the scale and complexity. Companies are still unsure how to combine multiple data sources to generate insights that are actually useful. Most companies are storing tens of terabytes of data in each warehouse with the capability to store petabytes. Pharmaceutical companies are cautious of the types of information they store because of the strict legal restrictions that are in place to protect privacy rights.

There is still strong interest in cloud storage if privacy concerns can be addressed. At least 27% of pharmaceutical companies are already using cloud storage solutions. The cloud is appealing to companies due to its low costs and ability to store the increasing amounts of data. As an example, Eli Lilly chose a cloud storage option because it would have taken 25 servers in-house. An option that may increasingly appeal to companies with privacy concerns is a hybrid solution of using both on-premise and cloud storage options. While only 10% of companies of companies cited this as the more attractive storage option, it holds a unique opportunity to store sensitive information in data warehouses and less sensitive information in the cloud.

## **Industry Trends**

### **Competition**

The pharmaceutical industry is highly competitive since each company is trying to bring similar drugs to market to treat diseases. Big Data has become a major factor in maintaining competitive advantage because it can help companies develop drugs and gain insight into diseases quicker. It is also a race to manage, process, and store the

data that holds the most insight so companies must continuously update processes, hardware, tools, and data sources.

### **HITECH Act**

Obama's HITECH Act provides hospitals with 5-7 million in incentives to switch over to electronic medical records by 2014. The HITECH Act essentially requires hospitals to use electronic medical records and failure to do so could result in 1 million worth of penalties. Not only does this help create cohesive medical records throughout the country, but it also provides pharmaceutical companies with access to detailed medical records to provide insight into patient health and drug interactions.

### **FDA Regulations**

The FDA has been instituting more stringent regulations for the licensing of safe and effective pharmaceuticals. This process requires significant time and money, especially if a compound does not get approved. Therefore, Big Data can help companies by eliminating unsafe and ineffective compounds earlier on in the research process. Therefore, companies can produce safer and more effective drugs and bring them to market sooner by analyzing Big Data.

### **Applications**

#### **Research & Development**

While pharmaceutical companies utilize Big Data for a multitude of internal processes, research and development has become a priority with the influx of patient data. At least 58% of Fortune 500 pharmaceutical companies are primarily using Big Data for research & development initiatives. The primary advantage of using Big Data is that it allows companies to make decisions more quickly and better target what drugs to develop and how to develop them. Only 10-12% of new drugs make it from the early phases of compound discovery to the market. Companies spend an average of \$4 billion and 12 years to bring a new drug to the market. Therefore, companies can

increase the number of successful drugs developed while decreasing time and money spent by using Big Data.

Big Data can accelerate the drug development process by facilitating biomarkers, which are characteristics of a pathogen, and using them to either validate or terminate compounds earlier in the research process. An example of this is Eli Lilly's Phenotypic Drug Discovery Initiative. Eli Lilly enables external researchers to submit compounds for screening using Eli Lilly's proprietary tools and data to identify if the compound is a potential drug candidate. Researchers can keep the intellectual property while it allows Eli Lilly to have a first look at potential compounds. Eli Lilly is able to find potential compounds quicker and bring new drugs to market while also benefiting researcher's initiatives.

Pharmaceutical companies can also use Big Data to improve insights into drug effectiveness and develop comprehensive risk-benefit profiles of patients. This can be done through a combination of health records, genomic data, and imaging data. Patient data can allow disease progression and medication effectiveness to be monitored over a patient's lifetime. By utilizing multiple types of patient data in conjunction with one another, a comprehensive patient profile can be used in predicting disease, evaluating the most effective drugs and therapies, and eventually lead to disease prevention. This will allow pharmaceutical companies to work alongside hospitals to create highly personalized medicine by understanding how different drugs interact with different patient profiles. Some pharmaceutical companies are already putting personalized medicine into action by partnering with hospitals to create local concept of care initiatives. Eventually, personalized medicine will be utilized on a large scale with the help of Big Data.

## **Clinical Trials**

Pharmaceutical companies are using Real World Datasets to improve patient outcomes through clinical trials. More than 80% of clinical trials in the United States fail to meet enrollment timelines and are stopped early because they cannot enroll or retain enough

patients to develop sufficient data sets. The majority of clinical trials do not make it past phase 2. Therefore, there is a significant amount of time and money wasted on unsuccessful clinical trials. Pharmaceutical companies can use both patient data and historical clinical trial data to improve outcomes and bring more effective drugs to the market.

Pharmaceutical companies are using patient data to match the right patient to the right clinical trial. One of the major reasons that clinical trials are unsuccessful is because patients drop out or do not comply with protocols. By using patient data, clinicians can help ensure retention rates, compliance, and relevancy of patients to ensure that the right people are in the study to generate the most successful trial outcomes. This can also help clinicians improve clinical trial design and protocol and reduce the time required to recruit investigators and patients.

Historical data from previous clinical trials can be used to identify adverse effects of compounds more quickly. Even if a trial did not produce enough results to be considered useful, results from multiple trials can be combined to generate insights. There is also potential to analyze clinical trial data in real-time to keep up with the latest clinical research to adjust processes more quickly. One setback is that many companies only share a portion of their clinical trial data with competitors. However, if multiple companies can combine clinical trial data then drug effectiveness could be evaluated quicker and more efficiently.

## **Issues & Improvements**

### **Privacy**

Pharmaceutical companies must navigate a variety of issues that are inevitable with the nature of the data they are dealing with. The number one concern that Fortune 500 pharmaceutical companies have when implementing Big Data initiatives is privacy. 29% of pharmaceutical companies view privacy as a major concern. Patient health records are highly sensitive and patients must either opt-in to having their information shared or



the data must be de-identified. Even de-identified data must be managed and stored compliant to stringent legal guidelines. Approval and consent to use data is required from multiple sources to assure that these guidelines are met. This can become a major obstacle when pharmaceutical companies are trying to gain access to third-party data such as electronic medical records.

*"29% of pharmaceutical companies view privacy as a major concern."*

### **Data Complexity**

The second biggest concern for pharmaceutical companies is the complexity of Big Data with 22% citing this as a top issue. The types of data that pharmaceutical companies are using are some of the most complex data to process and analyze. This is because the majority of the data is unstructured and has multiple levels of attributes that need to be processed. Therefore, pharmaceutical companies face the issue of having internal talent and the proper tools to manage, process, and analyze the complex data. There is also an overwhelming scale of data available especially with the influx of patient records. Companies must put in place processes to deal with the large amount of data to be able to properly store and utilize it to provide the most impact.

### **Data Quality**

The third largest concern, with 21% of Fortune 500 pharmaceutical companies citing this as an issue, is data accuracy and quality. The lack of standard data formats and definitions across the pharmaceutical and healthcare industries make integrating data from multiple sources difficult and time-consuming. Also, companies must ensure that data is healthy and consistent to guarantee that it is of high quality. Without high quality data, companies cannot generate impactful insights from their Big Data initiatives.

### **Additional Concerns**

Additional concerns include assigning ownership, costs, and communicating the value of Big Data initiatives to upper management. Companies must assign ownership for data sets, data quality, and Big Data strategies. One issue with assigning ownership is

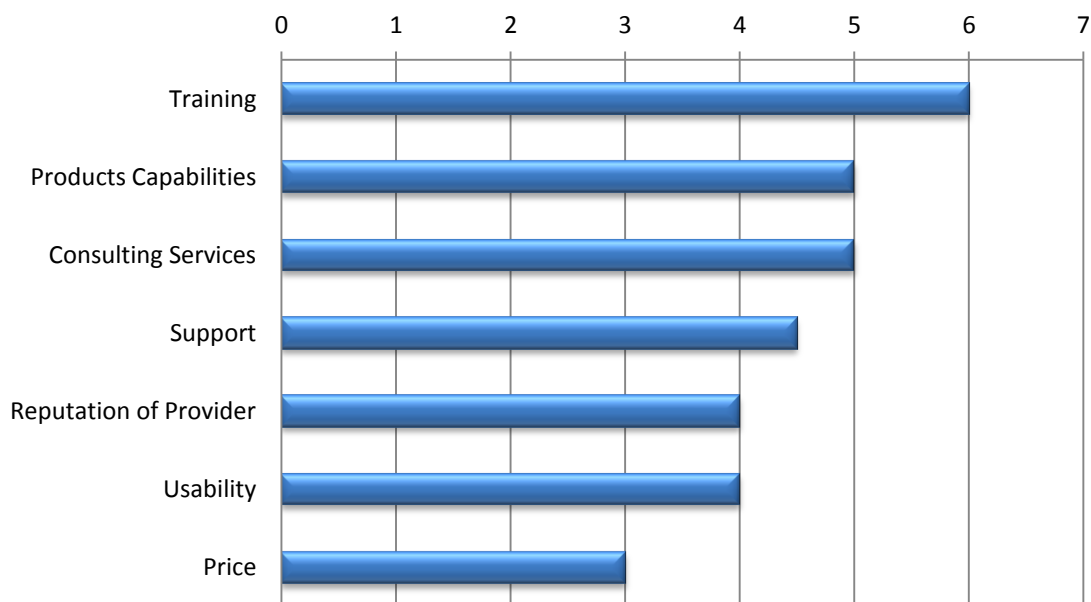
that with multiple Big Strategies throughout a company, it is difficult to decide who should own each one and if jurisdiction should fall to individual business unit leaders or IT leaders. Also, Pharmaceutical companies are concerned with the costs of storing the scale of data that is available especially with the privacy concerns of lower-cost cloud options.

## **Buying Process**

Most pharmaceutical companies have multiple Big Data strategies across departments. Individual strategies are owned by the individual business unit leaders. IT leaders also have a major part in each division's strategy because of their technical expertise. Buying power is held by individual business units and IT departments between the Director and Senior Vice President levels. The business units tend to have an idea of how they want to utilize Big Data for their department's strategy and the IT departments have a better idea of what type of Big Data tools they need to align with each initiative.

The budget for purchasing Big Data tools and platforms varies greatly between companies and departments. Business units tend to allocate a portion of their IT budgets towards Big Data purchases. It is estimated that companies spend anywhere between \$500,000 and \$10 million on Big Data initiatives.

Figure 1.1 addresses the elements which pharmaceutical companies believe are most important when deciding which Big Data tools to purchase. Each component is ranked on a scale of 1-7 with 1 being unimportant and 7 being essential.



**Figure 1.1**

## Partnerships

Pharmaceuticals companies partner with a variety of outside companies for storage, analytical tools, and third-party data. Common partners include:

Overall Solutions Providers	Industry Specific Solutions Providers	Third-Party Data Providers
<ul style="list-style-type: none"> <li>• Teradata</li> <li>• SAS</li> <li>• Oracle</li> <li>• IBM</li> <li>• Microsoft</li> <li>• Amazon</li> <li>• Cisco</li> </ul>	<ul style="list-style-type: none"> <li>• IMS Health</li> <li>• GNS Healthcare</li> <li>• Veeva</li> </ul>	<ul style="list-style-type: none"> <li>• Medco</li> <li>• Humana</li> <li>• Humedica</li> <li>• Quintiles</li> <li>• Qiagen</li> </ul>