Integrating Big Data for Health & Life Sciences Companies

Understanding the Full Value of Life Sciences Data Requires Integration

Across the life sciences landscape, organizations are overflowing with data: genome sequences, clinical trial data, electronic medical records, claims data, wearable devices, and quantified-self data. Competitive advantage awaits those that can extract value from all their data. Integrating this data, however, is a huge challenge. Companies must tackle an incredible volume, variety, and velocity of data and attempt to weave these pieces together to produce digestible insights for their business.

The explosion of data offers an opportunity for life sciences organizations to improve their competitiveness and market agility by looking at their data in new and complex ways—and insights need to be accessible through fast, reliable, scalable, and economical means. Pharmaceutical companies, facing the challenge of burdensome costs and difficulty in early failure detection of clinical trials, can especially benefit from data integration approaches and strategies, which allow them to generate comprehensive insights to better inform clinical trial planning, forecast market trends, improve customer relations, and improve business decisions.

Integrating big data sources is essential for creating a 360° view across all departments of an organization. Don't let an overflowing ocean of data drown you; use data integration to navigate your sources and understand your business in ways you didn't know were possible.

Life science organizations are finding value through data integration, the following examples illustrating just a small sample of benefits that can be achieved through this integration:

- **For Patients.** Enrolling patients in clinical trials based on more data sources; Automatically identifying patient cohorts with Electronic Medical Records; Leveraging genomic data for personalized medical compounds and products.
- **For Customers.** Understanding customer and potential customer sentiment; Tracking adherence of your customers.
- **For Companies and for the Market.** Forecasting market trends and customer need and using this information to speed up the drug development life-cycle; Automatically identifying key opinion leaders with social media data; Improving risk management by catching adverse events before they happen.

The volume, velocity, and variety of modern data sources in life sciences pose a unique set of challenges, summarized in the graphic at left. A 360° view of this data requires hybrid analytics approaches to account for contextual differences between structured and unstructured data. Storing these disparate data sources in a scalable way, integrating these data sources to facilitate rapid search and retrieval, and analyzing across these data sources to gain actionable insights pose a great challenge. The current methods to achieve these results won't cut it—new tools are needed.

The three Vs of Big Data, as they pertain to the life sciences. [1]

VOLUME

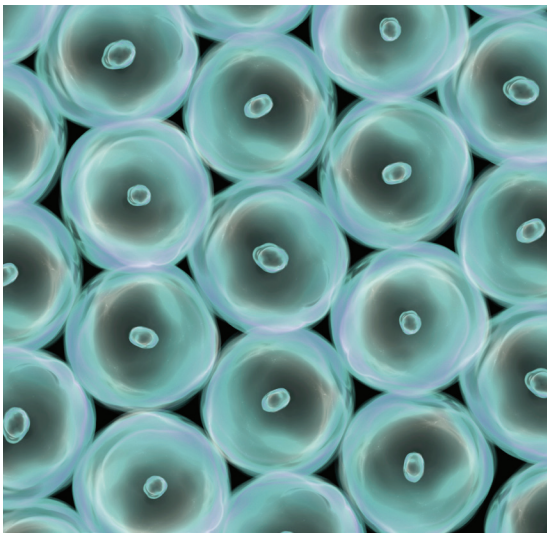
The sheer volume of genomic data (for example, the Illumina HiSeq-4000 can sequence 12 human genomes in 3.5 days, with a data output of 1.5 terabytes [2]) makes legacy data warehousing methods, data exchange, and classical statistical analysis inadequate.

VELOCITY

The velocity of data from wearable devices and cutting-edge smart sensors thrusts the field into the forefront of real-time streaming data processing and analytics. The metadata from these data sources must be catalogued in real-time, and the data normally must be transformed before it can be analyzed.

VARIETY

The wide variety of real world data sources unique to life science companies brings a new set of data challenges. Electronic medical records are often a hybrid combination of both structured and unstructured data, and social media data is especially unstructured and often extremely sparse.



New Data Integration Tools are Efficient and Useful

Relational database management systems have a proven record of enterprise deployments, but these legacy systems are not scalable, flexible, or fast enough to handle the variety, velocity, and volume of data in the modern era of life science organizations. Non-relational database technologies enable efficient integration and storage of structured, semi-structured, and unstructured data and easy exposure to advanced analytics platforms. (For an overview on non-relational, so-called NoSQL, design and features, please consult[3]) These database platforms are designed with the goals of scalability, availability, and high performance access in mind.

The first fundamental design goal of these new tools is scalability.
Cloud computing powered by the Hadoop Distributed File System (HDFS) enables an unlimited data capacity for rapidly growing data. Scalability is especially relevant to life science organizations because of the rising volume of genomic and medical imaging data and that over 70% of Real World Data comes from external sources.[4]

The second design goal of modern databases is availability.
Since data is stored on multiple distributed nodes, fail-safe mechanisms are built into the software, and can handle network failures and data corruption. Availability is ultimately such an important consideration due to the high costs of obtaining data from sources such as clinical trials and drug screenings and increasing requirements for ongoing analytics in real, or near-real time.

The final design goal is high performance access.
Data integration tools need to be built with large storage and computing intensive applications in mind, such that data can be written and read very quickly. As high-velocity data sources such as wearable devices are becoming more prevalent, high performance access becomes critical. Pre-sorting the data based on multiple fields—a process known as indexing—can make reading data even faster.

A data integration platform designed with the principles put forth by CARA offers you a 360° view of your data across your business, and the ability to seamlessly bring in outside data sources.

Efficient Data Integration for Marketplace Agility

Without efficient data integration strategies, an organization will not be able to leverage their data to make quantitative business decisions and remain competitive in the new data-driven landscape. It has been estimated that efficient data integration would reduce operational costs by as much as 20%.[5]

Booz Allen Hamilton's Cloud Analytics Reference Architecture (CARA[6]) offers a sound architectural design pattern for data integration and analysis that incorporates modularity, multiple use, and enables innovation:

- **Modular design.** Allows you to update or replace individual components of your platform as new technologies become available without having to redesign or rebuild the entire technology stack. For example, you can integrate a novel graph database technology and ingest your existing data easily without having to redesign the platform, and you can try out a new search strategy by inserting a novel indexing technology.

- **Multiple use design.** Offers the flexibility to connect different front ends and analytics tools to the platform in order to address multiple business questions, without the need to re-ingest or re-integrate data. Similarly, you can create different views of the data already present, allowing you to ask different types of questions more easily. This positions IT to provide utility to many different functions, using the same data pool.

- **Enabling Innovation.** Develop novel proof-of-concepts quickly through API access at all levels of the stack to enable low-effort and low cost novel analytics approaches to solutions.

A data integration platform designed with the principles put forth by CARA offers you a 360° view of your data across your business, and the ability to seamlessly bring in outside data sources. By doing so, you can further expose and analyze the data, rapidly producing quantitative answers to your business questions.

To showcase Booz Allen's CARA design principles, we built a demonstration data integration platform that leverages public data sources within the life sciences space. The life sciences demonstration data integration platform is hosted on the Amazon Web Services Elastic Compute Cloud (AWS EC2), where the public datasets are stored in their native format. The data is indexed with Elasticsearch to enable fast search and retrieval. We used Shiny by RStudio for the user interface and its extensive data visualization and analytics capabilities. Data sources in the demo are presented in the sidebar on the left.

Data Sources Integrated into the Platform Demo

DailyMed

http://dailymed.nlm.nih.gov/dailymed/
FDA-approved compounds. Data includes drug name, compound name, active ingredients and intended indications.

PubChem

https://pubchem.ncbi.nlm.nih.gov/
NIH-hosted compound database. Provides full chemical & physical properties of a compound, and is useful for compound synonyms.

Clinical Trials

https://clinicaltrials.gov/
US Clinical Trial database, which provides full clinical trial information for compounds and indications.

FAERS

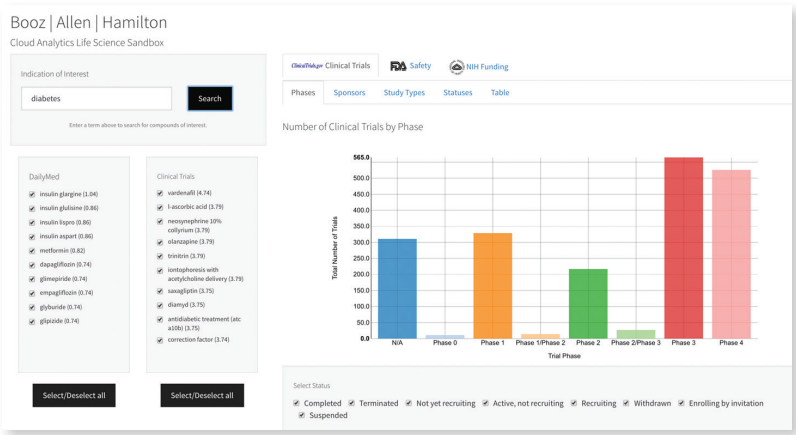
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm) FDA Adverse Events Reports for compounds. Provides both compound reaction information and patient outcome information.

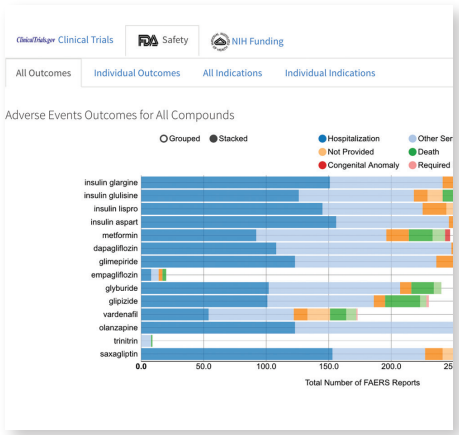
KEGG

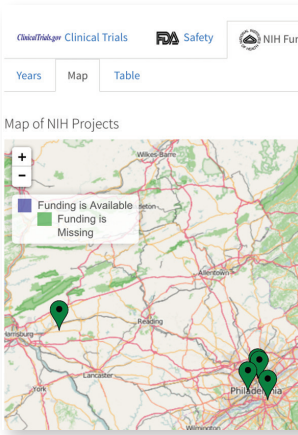
http://www.genome.jp/kegg/
The Kyoto Encyclopedia of Genes and Genomes is one of the most complete and widely used databases containing metabolic pathways from a wide variety of organisms. This data is useful to find connections to other diseases and compounds.

NIH RePORT

http://report.nih.gov/
Repository of NIH-funded research projects.







Above
The life science integration platform dashboard. A user can search indications (such as 'Diabetes') and receive a list of matched compounds. The first of three tabs are shown, which visualizes the associated clinical trial data.

Middle
The second major tab in the dashboard, which visualizes adverse events for the returned compounds.

Right
The third major tab in the dashboard, which provides a geo-tagged map of funded NIH projects and a NIH funding profile over time.

A user can access this platform to search for medical indications (such as "diabetes") and receive a list of compounds designed for the indication, as well as gain useful insights provided by the public data sources listed above, all in one user-friendly dashboard. The dashboard provides a clinical trials tab, which visualizes the breakdown of the total number of clinical trials for the returned compounds by clinical trial phase, sponsor, or status. You can even inspect the raw clinical trial data. The dashboard also provides an adverse events tab, where you can gain insights into the adverse event outcomes for the returned compounds, or drill down into the adverse events for specific indications. Lastly, there is a NIH funding tab, which provides a geo-tagged map of NIH-funded projects for the returned compounds, and a visualization of the NIH funding profile for the returned compounds over time.

Unleash the Power of Advanced Analytics

The true power of data integration lies in its interconnectivity. When stakeholders have a 360° view across the entire enterprise, they are empowered to unleash the possibilities of advanced analytics, gaining an understanding of the full story: what happened, why it happened, and what will happen next, in effect, moving from the identification of correlations to root cause analysis, an important factor in making the data actionable. The power of advanced analytics is the power to transform your company into an agile, competitive organization driven by data.

To learn how Booz Allen Hamilton can help your business thrive, contact:

Robert Zambon
Senior Associate
zambon_robert@bah.com
Tel +1 240-314-5654

Kelly Stepno
Senior Associate
stepno_kelly@bah.com
Tel +1 703-377-7184

Brian Keller
Chief Technologist
keller_brian@bah.com
Tel +1 816-582-2168

www.boozallen.com/commercial

References

1) Zikopoulos PC, Eaton C, deRoos D, Deutsch T, Lapis G. Understanding Big Data. New York: McGraw Hill; 2012

2) Illumina HiSeq-4000 specification sheet (http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/hiseq-3000-4000-specification-sheet-770-2014-057.pdf)

3) Wu, Stephen. An Overview of Present NoSQL Solutions and Features. Bloomington, IN. Indiana University (http://grids.ucs.indiana.edu/ptliupages/publications/An%20Overview%20of%20Present%20NoSQL%20Solutions%20and%20Features%20revised.pdf)

4) Realto life sciences survey. (http://www.reltio.com/about/news/2015/6/new-reltio-study-reveals-more-than-half-of-all-life-sciences-companies-now-data-driven)

5) Informatica White Paper. Big Data for the Pharmaceutical Industry (https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/executive-brief/big-data-pharmaceutical-industry_ebook_2341.pdf)

6) Escavarage, J., Guerra, P., Sullivan, J. Harnessing Big Data to Solve Complex Problems: The Cloud Analytics Reference Architecture. Booz Allen Hamilton. http://www.boozallen.com/media/file/the-cloud-analytics-reference-architecture-vp.pdf

Booz | Allen | Hamilton

Booz Allen Hamilton has been at the forefront of strategy and technology for more than 100 years. Today, the firm provides management and technology consulting and engineering services to leading Fortune 500 corporations, governments, and not-for-profits across the globe. Booz Allen partners with public and private sector clients to solve their most difficult challenges through a combination of consulting, analytics, mission operations, technology, systems delivery, cybersecurity, engineering, and innovation expertise. With international headquarters in McLean, Virginia, the firm employs about 22,600 people globally, and had revenue of $5.41 billion for the 12 months ended March 31, 2016. To learn more, visit www.boozallen.com. (NYSE: BAH)