**The Research**

1. [View the Apache Hadoop website](http://hadoop.apache.org/) - a free and open source project for distributed processing of large datasets.
2. Pick one project from the list on that page where it says "Other Hadoop-related projects at Apache". Read those materials.
3. Find and read materials on at least two (2) other sites that relate to Hadoop (not the Apache site).

While reading these items, think about how you might use that technology with a real-world data science problem.

**Your Response**

In no more than 250 words, write up your ideas about that use case and that technology, and include in your writeup the additional URLs which you read. Type your response in the text field and submit it to the dropbox.

Apache Hive is a piece of data warehousing software that is part of the overall Apache Hadoop ecosystem. It allows for working with large datasets in distributed storage which can be queried using SQL syntax.

One very modern use-case for Hive (and Hadoop in general) is in the pharmaceutical/biotechnology industry. As more drugs are discovered, subsequent drug discoveries become harder as niche spaces are filled in. In order to increase the speed of drug discovery, research and development groups can use big data warehouses to store and query all previous work in one location. Models can then be generated, and new ideas researched in a much faster and more holistic capacity.

(<https://mapr.com/resources/big-data-and-hadoop-for-the-pharmaceutical-industry/>)

(Also, I am not sure what technology they used on the back end, but this is just an amazing story: <https://www.fiercebiotech.com/biotech/deep-genomics-reveals-its-program-first-ai-discovered-drug-candidate>)

In the overall healthcare industry, data is being generated at a seriously unprecedented rate. In 2011, it was estimated that the US healthcare industry generated **150 exabytes** of data, and that was 8 years ago! It certainly doesn’t make sense to store that much data on a bunch of HDD in the server room of MGH, so Hive would provide a much cleaner and more robust way to store it across many distributed servers.

(<https://www.dezyre.com/article/5-healthcare-applications-of-hadoop-and-big-data/85>)