The term *data lake* emerged to represent the ability to store large amounts of data in a low-cost data storage architecture such as Hadoop. Then, once stored, the data could be “fished out” as needed for various purposes. For this activity, you will research big data architectures that enable data lakes.

An example of this can be seen in data regarding COVID-19. In reference to COVID-19, imagine how much data are tracked and stored as big data on COVID-19 vaccines.

[This article discusses the process of COVID-19 vaccines data](https://www.usatoday.com/in-depth/news/2020/12/16/covid-19-vaccine-data-supply-chain-software-immunization-registry-tiberius/3879655001/).

* Based on the article and what you have learned about data lakes, data management, and data collection, what do you think a data lake for COVID-19 vaccine data looks like?
* Share a security issue that may arise from all the nationwide data.
* What could be done to help secure the COVID-19 data lake?

<https://www.defense.gov/News/News-Stories/Article/Article/2446061/tiberius-platform-aids-covid-19-logistics-delivery/>

**Data lakes, data management, and data collection in the context of the COVID-19 vaccine:**

* There are **over 100 different software systems** all interacting to coordinate vaccine supply and demand
* Some software systems are **new**, built for the specific covid-19 vaccine effort:
  + **Tiberius** to keep track of the data from the data lake and distributer data like FedEx. It is the analysis engine that coordinates where, when and how many vaccines need to be sent out by “combin[ing] logistics information with census data to coordinate distribution of the vaccine” (Bajak & Loehrke, 2020).
  + **Vaccine Administration Management System** tracks when individual vaccines are administered, patient appointments, when the second dose is needed, etc.,
  + **Covid-19 Data Lake** where all the data is pooled
* Other software are **adapted** from health care and other government tracking systems
  + **State Immunization Information Systems** for keeping track of who has had what immunizations
  + **VTrckS** for vaccine orders
  + **Vaccine Finder** to show vaccine sites and their supply levels on a map. It was originally for the H1N1 pandemic in 2009.
* For such a data management system to work, it must have the ability for many **different people/groups to see data** (states, federal government, vaccine sites, clinics, etc.)
* **Personal information**, which is necessary in some systems like the Vaccine Administration Management System, which contains personal patient information so they can get the correct vaccines at the correct times, needs to be scrubbed when that data is sent into the data lake, which is more about high-level vaccine coordination.
* The data lake and the data management system (Tiberius) were **designed with specific goals and functionality in mind**. The data lake pulls in specific information to have the data needed for big picture analysis of what is happening/should happen with vaccine distribution. Tiberius automates certain analysis such as calculating weekly vaccine allotments to different entities.

**Potential Architecture of the data lake for the COVID-19 vaccine:**

* **Flexible cloud storage.** Because there are so many sources for the data lake, it needs significant storage space. I would imagine cloud storage would be the most feasible, although this is a federal government project, so they may have created dedicated servers to handle all the data.
* **Pub/Sub architecture**. Data is continuously being added: every time a shipment goes out, every time a shipment arrives, every time a vaccine is administered, inventory at every vaccine site, etc. I would imagine **continuous data streams** are arriving, potentially in a pub/sub architecture as publishers of information to different topics. Tiberius would be a subscriber to the topics to analyze and derive insights.

**A security issue** that may arise from all the nationwide data **is unauthorized access.** The article mentioned that over 600 representatives of 64 different jurisdictions have login credentials to Tiberius, the management software (Bajak & Loehrke, 2020). If someone was lax about their password or intentionally shared their login information with an unauthorized user, some sensitive data could be exposed. The article also said all personal information was scrubbed before entering the data lake, so sensitive data would be along the lines of potential scandals or coverups: who is receiving more/less vaccine supply or where there may be a shortage or crisis. Unauthorized users could also input false information, such as inaccurate inventory.

A way to help **secure** the data lake is to **encrypt** the data while it is in the lake and while it is in transit**.**

**References**

Bajak, A., & Loehrke, J. (2020, December 16). *Where’s the COVID-19 vaccine? Who’s been vaccinated? Here’s how we’ll know.* USA TODAY. https://www.usatoday.com/in-depth/news/2020/12/16/covid-19-vaccine-data-supply-chain-software-immunization-registry-tiberius/3879655001/