**Big Data Guided Research Task**

**What is big data?**

* Where is it coming from?
* What enabled it to emerge as a technology?
* What is it used for?

The Vs of big data - Depending on who you ask, there are 3, 4, 5 or 6 Vs of Big Data:

* Volume – 100s of Terabytes, or Petabyte range
* Variety – Handling tweets, emails, video, sensor data etc…
* Velocity – Not just large volumes of data, but large volumes of data incoming during a small time period
* Veracity – The data coming in are often raw, unfiltered, uncleaned and discrepancies need to be dealt with
* Variability – How often the structure or meaning of your data changes
* Value – It adds value to the business

How do you solve a problem like big data?

* Distributed systems (before Hadoop)

**Hadoop**

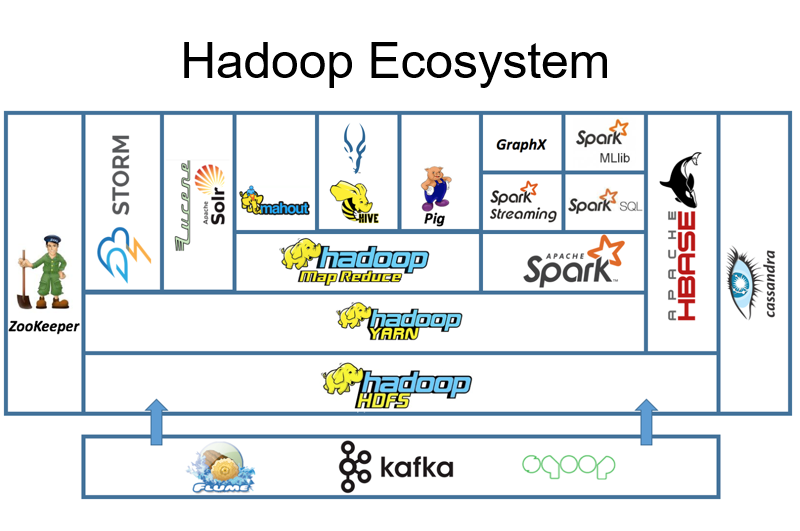
Hadoop was designed to address the downsides of distributed computing, and handles fault tolerance and programming complexity for us

What is Hadoop?

* When and why was it created

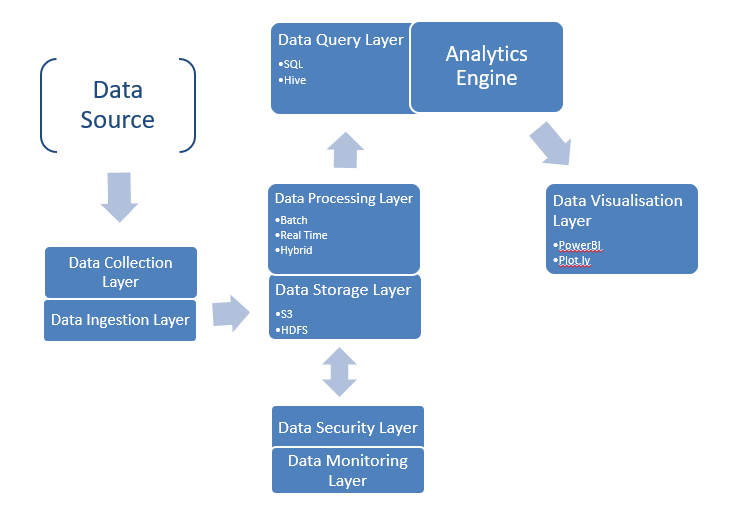
Main Components

* HDFS
* YARN
* MapReduce



Task:

1. Research the above topics in more detail, creating notes on the different points.
2. Create a diagram similar to this one, with the main Hadoop technologies added on, some have been done for you as an example:



This diagram shows how various layers and technologies involved within Big Data fit together. Data Engineers will mostly be concerned with Collection, Ingestion, Storage and Processing. But, the other way of thinking about the job of the Data Engineer is that they’re responsible for the arrows – the flow of data from one layer to the next.

Some tools will fit onto multiple different layers, think about what their primary purpose is, and if they do multiple things, you may need to account for that on the diagram.

1. Create a free community account with DataBricks by visiting: <https://databricks.com/try-databricks>   
   Important: when it asks you what cloud provider to use (Amazon, Microsoft etc), there is a small grey link at the bottom that you can click to use the community edition of Data Bricks

Log in to DataBricks at: <https://community.cloud.databricks.com>

Useful links to get you started:

<https://www.oracle.com/uk/big-data/what-is-big-data/#:~:text=is%20Big%20Data%3F-,Big%20data%20defined,especially%20from%20new%20data%20sources>.

<https://cloud.google.com/learn/what-is-big-data>

<https://hadoop.apache.org/>

<https://www.databricks.com/glossary/hadoop>

<https://www.ibm.com/blog/hadoop-vs-spark/>