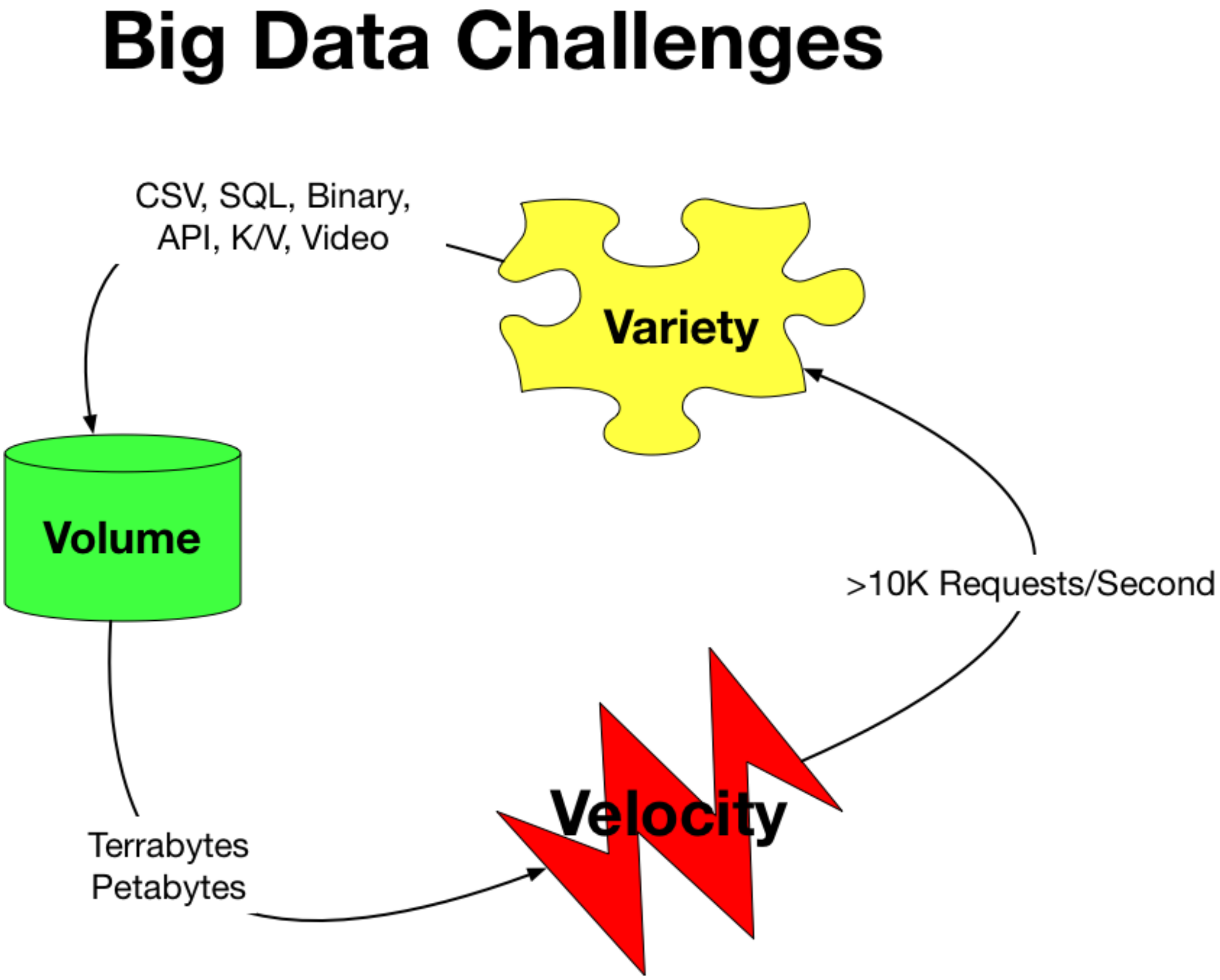

Big Data Challenges

Learn the three V's of Big Data is in the following screencast.

Video Link: https://www.youtube.com/watch?v=qXBcDqSy5GY[282]

Variety

Dealing with many types of data is a massive challenge in Big Data. Here are some examples of the types of files dealt with in a Big Data problem.

- Unstructured text
- CSV files
- binary files
- big data files: Apache Parquet
- Database files
- SQL data

[282] https://www.youtube.com/watch?v=qXBcDqSy5GY

Velocity

Another critical problem in Big Data is the velocity of the data. Some questions to include the following examples. Are data streams written at 10's of thousands of records per second? Are there many streams of data written at once? Does the velocity of the data cause performance problems on the nodes collecting the data?

Volume

Is the actual size of the data more extensive than what a workstation can handle? Perhaps your laptop cannot load a CSV file into the Python `pandas` package. This problem could be Big Data, i.e., it doesn't work on your laptop. One Petabyte is Big Data, and 100 GB could be big data depending on its processing.

Batch vs. Streaming Data and Machine Learning

One critical technical concern is Batch data versus Stream data. If data processing occurs in a Batch job, it is much easier to architect and debug Data Engineering solutions. If the data is streaming, it increases the complexity of architecting a Data Engineering solution and limits its approaches.

Impact on ML Pipeline

One aspect of Batch vs. Stream is that there is more control of model training in batch (can decide when to retrain). On the other hand, continuously retraining the model could provide better prediction results or worse results. For example, did the input stream suddenly get more users or fewer users? How does an A/B testing scenario work?

Batch

What are the characteristics of Batch data?

* Data is batched at intervals
* Simplest approach to creating predictions
* Many Services on AWS Capable of Batch Processing including, AWS Glue, AWS Data Pipeline, AWS Batch, and EMR.

Streaming

What are the characteristics of Streaming data?

- Continuously polled or pushed
- More complex method of prediction
- Many Services on AWS Capable of Streaming, including Kinesis, IoT, and Spark EMR.