Data source:

<https://dumps.wikimedia.org/other/pageviews/>

File structures:

Each file is an hour for every hour of a month within a directory of the month within the directory of the year. Each file uncompressed is 138 MB. The data goes back to 2015. A back of the envelope calculation of filesize \* hours in a month \* number of months= 229\* 730 \* 31=5,182,270 MB or 5060GB/ 4.94 TB.

The file is named according to its day and hour.

The example file I downloaded has 4,779,093 lines and there are 42,000,000 articles. This suggests that only pages that were viewed in the given hour are recorded. So if we were to build a database of sorts where each entry is a row that has timestamp, language and pageviews and we can assume that each row is 1000 bytes, the final database would be approximately 42GB.

The file looks like this:

en Barabási–Albert\_model 2 0

en Barachiel 5 0

en Barachois,\_Quebec 1 0

en Barachois\_(band) 1 0

en Barachois\_Pond\_Provincial\_Park 1 0

en Barack\_(disambiguation) 1 0

en Barack\_Obama 296 0

en Barack\_Obama's\_farewell\_address 1 0

en Barack\_Obama,\_Sr 1 0

en Barack\_Obama,\_Sr. 5 0

en Barack\_Obama\_"Hope"\_poster 10 0

en Barack\_Obama\_"Joker"\_poster 1 0

en Barack\_Obama\_Academy\_of\_International\_Studies\_6-12 3 0

en Barack\_Obama\_Democratic\_Club\_of\_Upper\_Manhattan 1 0

en Barack\_Obama\_Presidential\_Center 18 0

en Barack\_Obama\_Sr 4 0

en Barack\_Obama\_Sr. 22 0

It is space delimited, the first column is the language, the second column is the article title, the 3rd column is the number of page views in the hour and I don’t know what the 4th column is.

Information on the page views data is here:

<https://meta.wikimedia.org/wiki/Research:Page_view>

Our mission is to download the ~3.1 TB of data, create some sort of analogue to an SQL table that has the columns of time stamp, language, article and pageviews where we can sum views by date or just views in general. We then need to do some sort of analysis. Some example questions to answer:

What are the pages that are frequently visited often?

What are the top pages over our analysis period?

How many page views per language per time?

Jordan’s proposal for system to use:

A combination of Cassandra and Spark.

<https://opencredo.com/data-analytics-using-cassandra-and-spark/>

We should build a fail proof system . See

<https://opensource.com/life/16/5/basics-cassandra-and-spark-data-processing>

“Cassandra has the ability to be always on in spite of massive hardware and network failures by utilizing a design first widely discussed in [the Dynamo paper from Amazon](http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf). By using a peer to peer model, with no single point of failure, we can survive rack failure and even complete network partitions. We can deal with an entire data center failure without impacting our customer's experience. A distributed system that plans for failure is a properly planned distributed system, because frankly, failures are just going to happen. With Cassandra, we accept that cruel fact of life, and bake it into the database's architecture and functionality.”

7/29/2017 Jordan

<https://stackoverflow.com/questions/4775388/how-much-data-per-node-in-cassandra-cluster>

“

|  |  |
| --- | --- |
| *9down voteaccepted* | *1 TB is a reasonable limit on how much data a single node can handle, but in reality, a node is not at all limited by the size of the data, only the rate of operations.*  *A node might have only 80 GB of data on it, but if you absolutely pound it with random reads and it doesn't have a lot of RAM, it might not even be able to handle that number of requests at a reasonable rate. Similarly, a node might have 10 TB of data, but if you rarely read from it, or you have a small portion of your data that is hot (so that it can be effectively cached), it will do just fine.*  *Compaction certainly is an issue to be aware of when you have a large amount of data on one node, but there are a few things to keep in mind:* |