Distributed Systems Architecture

brought to you by Alexey Grishchenko

The Myth of “Unstructured Data”

Hadoop is known to be an ideal engine for processing unstructured data. But wait, what do you really mean by “unstructured data”? Can anything be considered as a “data” if it does not have a structure? Let’s start by taking a look at the historical brief.

First computers have appeared as a complex calculation engines. They didn’t have a permanent storage and were used mostly as calculators for scientific computations. But very soon their creators realized that some computations and some algorithms require additional information in order to work, and this information should be stored somewhere. They introduced the first storage devices. From the very beginning the data processed by computers was strictly structured. Each bit of the space mattered for these complex engines taking many floors in datacenters of that time. And this is why complex data structures and compression algorithms were introduced, to better structure and compact both the code and the data.

The age of the databases started a couple of decades after this. The amount of data stored by computers has increased. At specific point in time it started to be reasonable to introduce an engine that would manage data structure and give you an interface for the data manipulation. The most popular concept for the data storage and processing was a relational one – the working set consists of “relations”, and each relation is an unordered collection of unique “observations”. But it was not the only approach – there were also object stores, graph databases, spatial databases and so on.

In general, databases processed big amounts of numbers with some text metadata. When the storage became cheaper, new data types were introduced to accommodate the increasing end user requirements.

New data types were able to handle large text fields (like Oracle CLOB, PostgreSQL varchar TOAST) and large binary objects (like Oracle BLOB, PostgreSQL bytea with TOAST), which made it efficient to store big objects in a database. This opened the doors for the next advancement, introduction of the XML, JSON and other data formats as a database objects.

So here we come to the subject of this writing. What is “unstructured data”? From the creation of the computer till nowadays all the data computers worked with was completely structured. Imagine a text file. It has a file name. It lies on the file system of your computer. It has creation date. It has a specific size in bytes. And of course it has the text contents itself. Should we consider this information as unstructured? I don’t think so. If we need to store text files we can perfectly use RDBMS for this purpose, creating a table like “*filecollection (fileid bigint, filename varchar, filetype varchar, created timestamp, modified timestamp, size bigint, owner bigint, contents varchar)*” and process it by the RDBMS.

Now you can say that you cannot easily extract the knowledge from this text and this is why it is unstructured. For instance, by this you are considering the case with text analytics over blog posts. But imagine a simpler case. You have a grocery shop with its database containing a perfectly relational data: product names, product prices, amount of products left in stock, customer purchases information and so on. This is clearly structured data, isn’t it? But does this data allow you to extract the knowledge from it? Which products are most likely bought together, or how the seasoning affects customer purchases? No, you cannot do it without special machine learning or statistical algorithms. And it is similar for the text – you have some meta information about its contents, but to be able to perform sentiment analysis and name entity extraction you have to apply specific statistical and ML algorithms.

This is the same for images and video. JPG file your iPhone has produced is not only an image. It contains lots of metadata about the image including the GPS coordinates of the place you’ve taken this picture at. Video files are pretty the same. Again, you can easily extract some metadata from them but if you want an in-depth analysis you have to apply specific algorithms.

So why do you grocery shop data “structured” and free text/image/audio/video data “unstructured”? There’s completely no use in it in my opinion.

What about JSON and XML? The modern way to call them is “semi-structured” data. Why the hell “semi”? Both this formats are completely defining the structure of your data, and each of them can be stored in RDBMS either as a text or represented as a set of relations if your database does not support nested objects, maps and arrays (while some databases have a support even for this kind of things). Some databases even created special container types for these data structures, like jsonb in PostgreSQL. This data is perfectly structured, these XML/JSON are just specific serialization formats that were introduced to make it easier to transfer the data between different applications, i.e. make this data human-readable. No one in their mind would store raw JSON collection in their Java application: for each JSON exchange format they would create a class that is capable of being serialized as JSON and deserialized from JSON. And it is the same for XML, which is the base for many formats like HTML and SOAP.

According to the Wikipedia, unstructured data refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Thinking about this, how can the information not have a pre-defined data model? Could you have at least one example of this? Logs are usually considered to be unstructured data. But wait, for any of the tools generating logs you can find a full description of the format that this tool uses for logging the data, plus many of them are using log4j, which defines a basic structure of your log. What about images? What structure can an image have? Each image has the mandatory metadata (width and height in pixels for instance) and optional one. If the image is considered to be “unstructured” data then how does the image editor process it? In fact, images are perfectly structured and the image format completely defines the structure of the image file. So coming to think about this, the only example of the “unstructured” data can be a white noise – it does not have any structure, as it is completely random. But it does not make much sense to store write noise. And even if you do, you would store it as a file with metadata (filename, file size, white noise generation algorithm used, algorithm parameters, etc.), which in fact make this data structured.

The terminology has to be clear. There is no such a thing as “unstructured data”. But there are a number of different things: free text data, image data, audio data, video data. All of them have much in similar. To make use of them you have to apply complex algorithms for the knowledge extraction, but this fact does not make this data unstructured, it just makes it hard to interpret by computer.

And here goes the main question: how is Hadoop related to the problem of “unstructured data”? The correct answer is: it is completely unrelated. Hadoop as a distributed processing engine gives you an option to process huge amounts of information on a distributed cluster without spending much efforts on the framework, allowing you to concentrate on the logic that processes each of the records (mappers) and aggregates the final results (reducers). Facebook stores 300PB of data in their Hadoop clusters, but is this data “unstructured”? Hell no, they are using

ORCfile format, and in this format each file represents a row-columnar table. What about Twitter with its 30+ PB of data? They introduced Parquet, row-columnar data storage format and again, each file represents one table (with a support of nested objects, arrays and maps). Then what about the famous “schema on read” approach? Imagine Twitter to store their information as a set of pure JSONs, how wasteful this would be – instead of 30+ PB their storage footprint would be at 100+ PB even given a good compression applied.

I agree that Hadoop is a good engine for processing full text/images/audio/video data. But you can take an advantage of Hadoop over pure storage only in one case, when you have to reprocess the full collection of objects you store from time to time. Imagine you’re a scientist working on facial recognition algorithms. You develop these algorithms locally and from time to time apply them to all the images you have collected over the internet, and run validation on top of the sample of this data. Or imagine you have an image search service. You are continuously improving your image search algorithms, which forces you to rebuild the search index over your image collection many times.

But these cases are not very typical. Usually enterprises are buying Hadoop to analyze mythical “unstructured” data that in fact they don’t have. Some of the CTOs even think that just buying Hadoop cluster would open them a way to extract the data from Facebook, Twitter, LinkedIn and other social networks, which in their opinion cannot be done without Hadoop because their data is unstructured. Don’t buy into it, this is a pure marketing.

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE

© marketoonist.com

The main idea why Hadoop plays well is its price/performance factor, which is much lower than any relational database and MPP database can offer. With Hadoop you can get a license at <$5k per node, while MPP would cost you 10x this amount. This way Hadoop opens you a way to store “less important” data and process it to extract smaller pieces of the value. Like storing and processing the logs of all your applications, storing and processing the data of social networks and so on. They contain a way less value than the sales numbers of your grocery store, but this data still can be used to improve your sales a bit. Be skeptical and remain professional at all times.

Share this:

✉ in 55 f 🐦 G+

This entry was posted in Enterprises, Hadoop and tagged analytics, big data, hadoop, marketing, unstructured data on July 28, 2015 [https://0x0fff.com/the-myth-of-unstructured-data/] .