One year on: Apache Spark continues a winning battle against data management misery

Take a deep breath and herald the arrival of Apache Spark – the missing link in data management set to make all our lives easier. I'll explain why Spark is so good shortly, but first let's get started by putting it into context.

Back in the old days of IBM DB2 and Oracle, we used to look at our data in very structured form such as tables and spread sheets. No surprise there, considering that tabular data and relational databases were the most efficient way to manage data with the computing power we had. Normalised, structured data spawned a very powerful ecosystem of patterns and tools to analyse data.

The days of structured #Data are done with says @ArtyomAstafurov CLICK TO TWEET

ALL GOOD STUFF YES?

NO.

These systems only worked as long as the data was structured and normalised, and the format of the data sources was deterministic. The old systems failed to accommodate for when the landscape dramatically changed, i.e. when we added things such as social feeds, news streams, and sensor data into the mix. Data that we now call unstructured data or more commonly, Big Data; in other words, data that couldn't be analysed with the tools and methods we inherited from the days when normalised and relational data prevailed.

What did we do about Big Data? The solution came from Google and Yahoo! In mid-2000 Google issued a paper on an algorithm called MapReduce, which streamlined analytics of Big Data, making it easy to run it on multiple parallel machines. A year later, Yahoo! released Hadoop which for many years became the industry standard for analysing Big Data.

So what issues do we face today?

Unfortunately for Hadoop and similar tools, technology never stands still. Big surprise there!

System configuration management has become a significant factor in how fast we are able to develop and test a new analytics algorithm and it is something we now have to adapt to.

ENTER APACHE SPARK

Enter Apache Spark. The timing couldn't have been more perfect. A project that started in 2009 in the AMP labs of UC Berkley has quickly evolved into a top level Apache project with over 465 contributors by 2014. We are now witnessing its genius as it has unprecedented positive affects on the data management industry.

In contrast to #Hadoop's implementation of #MapReduce, #Apache #Spark provides performance up to 100x faster CLICK TO TWEET

In contrast to Hadoop's implementation of MapReduce, Spark provides performance up to 100 times faster. By allowing user programs to load data into a cluster's memory and query it repeatedly, Spark is well-suited to machine learning algorithms. Spark allows us to interactively explore data.

Say goodbye to the days when we were forced to model data with one set of tools and then implement and run our models with another, and say hello to a more streamlined data management approach. Spark closes the gap between data discovery and running analytics in production, giving an all in one approach to looking at data by using state-of-the-art Functional Programming approach.

But hold your horses, because it gets even better. In the very near future Spark will continue to change and revolutionise the way we look at data, giving us the foundation to get the most out of it in a lean and agile manner.

Original article — http://www.comparethecloud.net/articles/one-year-on-apache-spark-continues-a-winning-battle-against-data-management-misery/