Hadoop World 2013 - Day 3

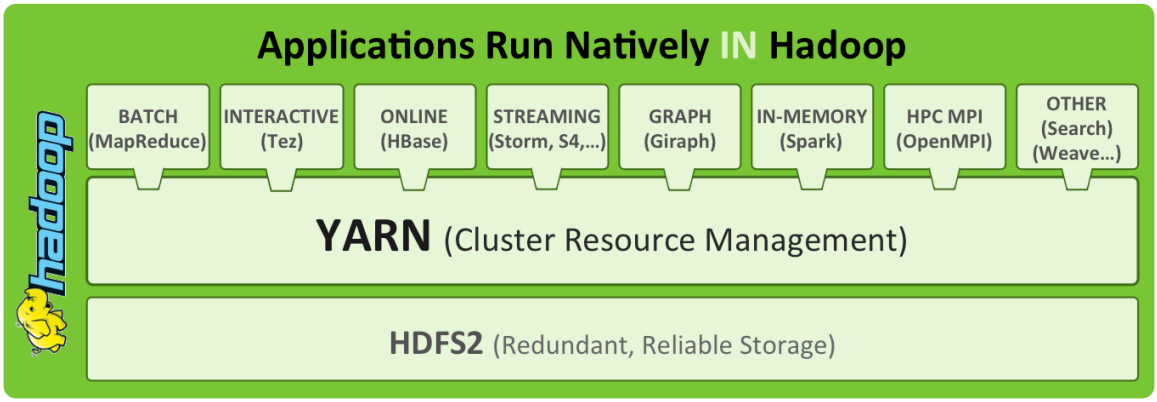
We made it! Like with Hadoop World 2013 - Day 1 and Hadoop World 2013 - Day 2, I wanted to offer up a recap of the final day and make sure everyone knows I'm more than excited to share my experiences, knowledge, and even my opinion for those who ask it. Seriously, what a great event during an exciting time for open source data fabric technologies.

The day started off like the previous with a rapid fire of 5-10 minutes talks during the traditional keynote window.

- Doug Cutting was up first. He said he was asked to share his predictions and like any true programmer stated that seeing the future was difficult. Again, like any good programmer he slowly walked us through his disclaimers and reasoning (he reminded us several time he can't see the future!). He was very humble about the beginnings of Hadoop and how much the community has grown it up from where they all started around 10 years ago. He finally got to his declaration that "**even transactions possible on Hadoop**". This is a giant statement and really fit in with Mike Olson's "enterprise data hub" direction presented the day before. It was further echoed in his richer prediction; "**it's inevitable that we'll see just about every kind of workload be moved to this platform; even OLTP**". If you buy into this line of thinking then you can't help but get excited by Doug's "**we are in the middle of a revolution in data processing; revolutions are scary times**" line and his summary quote of "**the future for data is Hadoop**".
- Josh Klahr from Pivotal was up next with his talk on being *data-driven*. He incorporated Marc Andreessen's "**every company needs to be a software company**" advice into his talk. He piggybacked much of Doug's thinking by sharing his belief that Hadoop is the "**next change in the 'data fabric'**" – he declared the only other two events at this level where the introduction of mainframes and then in the 80s the proliferation of RDBMS. Also like Doug, Josh's belief is that "**real-time apps are a critical evolution for Hadoop**".
- SAP's David Parker told us about their $10,000 competition they are running on their Big Data Geek Challenge (this is for the idea, not a full implementation).
- We heard some cool talks from Shawndra Hill, Will Marshall, and Jim Kaskade. My favorite quotes from Jim (Infochimps CEO) were "**big data starts with the application**" and his follow-on "**don't have a big data reference architecture, have some use cases**". We all should take this to heart as the real question for Equifax is how can Hadoop be useful to us – *not what version & vendor should we be using*.
- Peta Clarke & Donna Knutt came out next and gave us an overview (and call for help) on the Black Girls Code initiative. Good stuff – "Like" them at https://www.facebook.com/blackgirlscode.
- The last two speakers where an unlikely sounding pair, but they ended up same much more in common than either probably expected.
    - Douglas Merrill (the guy from All data is credit data...) gave a talk debunking the "myths of data science" (as well as mathematicians and statisticians – *of course, that's what he is*). He said he doesn't like the title Data Scientist and prefers Data Artist instead. He left us with "**we don't need more R, we need more artists**".
    - Douglas was followed by Foster Provost who is a professor from NYU and author of Data Science for Business. He admitted he like the title Data Scientist. He also seemed to debunk some of the current thinking that we get dramatically better predictive analytics when we analyze all the data, not just a rich sample. He says that often doesn't give him much additional lift. He states that the big increase in lift is when we add in additional varieties of data into the equation. Sounds like another good lesson for us to hear.

SIDEBAR: During Douglas Merrill's talk he dropped a joke that is still cracking me up. It is probably an old one, but it was the first time I heard it. *What is the difference between an introverted engineer and an extroverted one? One of them looks at his shoes when he is talking to you; the other looks at your shoes when he talks to you.*

Besides networking and meetings with vendors in the showcase pavilion area, the rest of the day followed the typical hour-long sessions. Here's a recap of the major ones I visited.



- As you could have guessed from the graphic above, I did attend the talk from Hortonworks on YARN where I captured the "**YARN is the OS of the Hadoop cluster**" quote that I actually buy into. Not trying to be controversial, but once could mentally walk this one through to its logical extension/conclusion, *but... I said I wouldn't!* We worked through the obligatory v1 vs v2 comparison/contrast slides and then spent the bulk of the time discussing YARN itself, apps on YARN (many of them still maturing along with YARN), YARN best practices should you want to add your own "OTHER" (see upper-right of above graphic), and outstanding work still to be completed on YARN. *I'll admit it...* ***I'm stoked about YARN!!***
- I then checked out another Hortonworks talk on the Stinger initiative to bring 100x (or more!) performance improvements to Hive (YARN to the rescue!). This talk was a bit more advanced for me as it really was a very detailed walkthru of some of the major changes they are doing (i.e. I'm not Hive expert). What was clear to understand was the TPC-DS queries they were showing improvements on (Query 27

had 190x improvements while Query 82 saw 200x).

- SAS' Paul Kent shared their approach of taking their compute engine into YARN (seeing a theme here?) to leverage Hadoop v2 to run their models in a massively parrallel way – all while leveraging Hadoop's data locality paradigm. Pretty promising stuff, but I'm sure EXPENSIVE!! Also, that data that will sit in HDFS will be in SAS format. To help a little with that, they allow you to read SAS formatted files via PIG. They've also devised a way for you to surface other data file types (via an interface to implement) in such a way to allow their YARN-based compute engine (technically a "YARN-compliant application") to read them. No real word on the pros/cons on that approach or too much detail, but it did sound pretty immature at this point.
- I'm realizing my "quick" blog post is rambilng on... Here's some thoughts on just a couple more sessions I attended in order to wrap this up.
    - Ravi Hubbly from Lockheed Martin gave a talk on migrating mainframe apps to Hadoop. He pointed out a number of similarities between the mainframe and Hadoop which included, file-based storage formats, big focus on batch, and my personal favorite; both approaches store flat files in whatever format they are in and then apply structure/schema at read time *(often COBOL copybooks for the mainframe, but EFX's "all kinds of weird file formats" really are exactly what the Hadoopers of the world have been doing – "it is what it was", or David's "it STILL is a good idea")*. The data and metadata migrations are actually pretty easy efforts (the presenter is clearly calling for vendors to help – copybook to popular HDFS file formats would be particularly useful) and I was glad to see he circled back around to the "elephant" *(get it?!?!)* in the room – the business logic! There is no magic answer for this, but he did state that in their experiences they have found that COBOL programmers adapted to Pig much faster than Java programmers did. I guess I need to try my hand at Pig Latin a bit more to offer an opinion on that one (of course, I'm one of those people who has been a professional COBOL and Java programmer).
    - Bobby Johnson from Interana gave us some best practices and advice on getting the most out of time-series data. He shared his experiences from his six year run as Facebook's Director of Engineering for Scaling & Performance. He also couldn't stop mentioning his Scribe project that I'm betting is getting a bit "long in the tooth". 

All in all, I'm glad I was able to participate and am hopeful I get the opportunity to do it again. I ensure all of our attendees (see Hadoop World 2013 - Day 1) are pulled back together for a retrospective to see what we collectively took from the conference as well as next steps. This may include upcoming Tech Forum; Equifax's Internal User Group presentations/demos. We will also see if any changes need to be made to the Equifax Position on Hadoop Distributions based on the information we have learned.