Hadoop World 2013 - Day 2

Wow! Day 2 has been great. Talk about drinking from a fire hose. It actually reminded me of the first time I went to JavaOne back in 2001. Just like then, there is a TON of excitement and buzz. Energizing for sure!!

First up was the keynote***S***. Yes, that's plural. Unlike many conferences where you get a focused talk from one person (or tight-nit team), today's kickoff was a series of 5-10 minute talks that felt a bit like being at a TED event. Here are some highlights.

- Mike Olson, Cloudera's Chairman & Chief Strategist, ran folks through the history of the Hadoop World conference. At the first event, 2009, there were 700 folks who came. This week's event was a sell-out that is bursting at the seams with 3000 folks running around. It is clear that Cloudera is ready to build upon the success of Hadoop being a great batch framework and take it to the next level by focusing on handling the real-time workloads their customers are screaming for. He briefly mentioned the upcoming release of CDH5 and was labeling it as an "Enterprise Data Hub" as he hated the phases data reservoir and data lake. Their areas of concentration are now security, data lineage/maintenance/governance, ensuring there can work with a "rich collection of processing engines" (real-time, including Impala of course, and batch), and connecting with the rest of the tools in use within enterprises. They are also working with DataBrick's Spark for real-time workloads. It was awesome to hear the belief in the "community" that Cloudera has. That's been a VERY common theme here this week and as a customer, I'm so glad to hear it. Mike was very gracious & humble about the fact that we all will have plenty of vendor choices to work with.
- The Strata Awards where then handed out. Winners included Facebook's Datacenter Visualizations, Captricity, and AADHAAR.
- Jack Norris, MapR's Chief Marketing Officer, worked on exposing various myths. First up was the "wrestling elephants" in which he said MapR, and all the other distros, come from a common lineage, but that some have expanded upon this for performance concerns. He's mostly right on this one, but let's move on. He was right which he compared/constrasted Hadoop vendors to NoSQL vendors who surely don't have any common approach, much less base open source project that all rally around. He did toss out HBase as one of the leaders because of its obvious coupling with Hadoop. He finished up with a bashing of ASF's HBase implementation by showing that MapR's M7 really is a seperate product than what everyone else is doing which basically negated his early comments of "building upon" instead of "rebuilding" the Apache projects.
- Ken Rudin, Director of Analytics for Facebook, explained that FB started with Hadoop and has augmented it with RDBMS technologies which is clearly backwards from most larger/older companies who have augemented the RDBMS environments with Hadoop. He was quite clear that the goal of analytics was not actionable insights, but that is was driving impact (ex: move a metric, change a product, or change a behavior or process). His "if nothing changes, you've made no impact" quote resonated with the audience.
- Tony Salvador from Intel gave an interesting talk on the "Personal Data Economy" that you can find out more about by visiting http://wethedata.org/.
- Quentin Clark, who is from Microsoft, had the best slide deck of the day (MS always has cool slides) and he directed us to http://microsoft.com/bigdata which interestingly enough routes you to a SQL Server page. Hmmmm... That said, their HDInsights "Hadoop on Azure" is actually Hortonwork's distribution as described here.
- Lots of other great talks from folks like Ben Werther; loved his "***BI = BS***" slide!

Then it was on the the normal conference sessions that one would expect. Here are some quick bits from the major sessions I attended.

- I first saw a talk on Google's Hadoop Anonymization Toolkit (HAT) that looked rather promising, but hasn't yet been released to open source.
- An engineer from Spotify walked us through some of the key roadblocks and issues his team faced when they grew their 60 node cluster to 690 nodes. It was half funny and half scary!!
- A few us then learned about Parquet; an open columnar storage for Hadoop that Cloudera is helping to flesh out (and leverage with Impala).
- I then made the mistake of attending What's Next for HBase– not because it wasn't solid content on the roadmap which included multi-tenancy, performance isolation & priority features, but because it let me know how little I actually know about HBase. 😉 I did find out that Facebook's "Messages" (I think that's the email peice, not the "wall posts" or the notifications) uses HBase. I also learned about salesforce.com's Phoenix JDBC "skin" over HBase that looked interesting. All in all, it was clear that HBase is getting a lot of "engineering love".
- It was reassuring to see that Apache Sentry is still coming along that is focused on securing the Apache Hadoop ecosystem.
- The true disappointing session of the day was about Microsoft's REEF. Knowing of their investment in Hortonworks and hearing their keynote speaker, I was actually optimistic when it sounded like they were going to be a true player in the emerging YARN space. I need to stay "positive" in my blog, but don't mind sharing some thoughts if you give me a call or swing by my desk. In a nutshell, they should call this Retainable Evaluation Execution Framework the ***YA-YA-RN***; the Yet Another Yet Another Resource Negotiator. When the presenter finished, you could have hear a pin drop and cut the tension in the air with a knife. If I wasn't so upset at what I just heard I would have actually felt sorry for the presenter. Again, if you want to hear more about this please reach out to me directly. 🙁

As with Hadoop World 2013 - Day 1, overall this was another winner and I'm looking forward to the last day of this short conference and to get back to work to share more of what I've learned.