ETL, ELT and Data Hub: Where Hadoop is the right fit ?

Posted by **datumengineering** on November 17, 2013

Few days back i have attended a good webinar conducted by Metascale on topic **"Are You Still Moving Data? Is ETL Still Relevant in the Era of Hadoop?"** This post is targeting this webinar.

In summary, this webinar had nicely explained about how enterprise can use Hadoop as a data hub along with the existing Datawarehouse set up. "Hadoop as a Data Hub" this line itself raised lot of questions in my mind:

1. When we project Hadoop as a Data-hub and same time maintain the datawarehouse as an another data (conventional) repository for the enterprise then won't it be creating another platform in silos? Presenter in the webcast repeatedly telling about keeping existing datawarehouse intact when developing Hadoop as a Data Hub. Difficult to digest :(
2. Next question that would arise is: challenges in Hadoop environment as Master Data Management and Data governance platform. I don't think Hadoop ecosystem is mature enough to swiftly handle the MDM complexity. As far as data governance is concerned Hadoop ecosystem lacks in applications which are required on top of Hadoop for robust data governance.
3. Why to put lot of energy to build compatibility of ETL tools like Informatica with HDFS to connect existing ETL infrastructure with Big Data? I feel this is a crazy idea. Because you are selling cost effective solution with some under the cover cost. Obviously, Informatica will not give you Hadoop connector as "Free". There are many other questions other than cost like performance, business logic stage etc.
4. Also there is a big bet on Hadoop to replace existing ETL/ELT framework to push transformation to Hadoop considering its Map Reduce framework. I partially get this idea long back. But, still not convinced when:

    - Your use case doesn't support Map Reduce framework during ETL.
    - You process relatively small amount of data using Hadoop. Hadoop is not meant for this and takes longer than it supposed to be.
    - You try to join some information with existing datawarehouse and unnecessary duplicate the information at HDFS as well as at conventional RDBMS.

Now, having these questions in place doesn't mean Hadoop can't be projected as a replacement/amendment of existing datawarehouse strategy. On contrary, I could see some different possibilities and ways for Hadoop to sneak-in into the existing enterprise data architecture. Here are my few cents:

1. Identify the areas where data come with once write and many reads. Most importantly, identify the nature of the read. Ask yourself that the read is straightforward or joined or aggregated? All such situations in case of BIG DATA can be efficiently handled on HDFS. If you

don't know in advance then data profiling and capacity planning will be decision maker here to identify whether this data should go to your RDBMS or HDFS. However remember, if your queries is more ad-hoc and you are planning to move it to HDFS then you need a skill more than Hive & PIG.

2. Use Hadoop's HDFS feature more than the Map Reduce. I mean distributed storage to minimize effort to back up and data replication. This will be cost effective in comparison to DBA costs. For example, archive data on HDFS than tap drives. So your data never retire for analysis. Entertain MR intelligently whenever you could see the opportunity to break down your calculations into different parts i.e. MAPs.
3. Identify the data which is small and can be fit into distributed cache in HDFS. Only this can have an entry into HDFS. However, rest of small (not BIG) data can stay on RDBMS. Again, **Capacity planning** is major role player.
4. **Now it comes to ETL**: I am really happy to see Hadoop & HDFS here. But not with Informatica, Data Stage or any other ETL tools (i don't know much about Pentaho). I must appreciate and support **Metascale webinar** . They have given a right approach to take Hadoop as an Extract, Load and Transform framework. Yes, this is the only way to do right transformation on Hadoop. Let's rename it to DATA INTEGRATION. The moment you start thinking about ETL tools it means you are taking your data out of Hadoop and the moment you take data out, you are going against all the purpose of using Hadoop as a data processing platform. Isn't it killing the idea of doing transformation on Hadoop using MR and also the idea of brining your overall cost effective down? However, I'll be open to learn the right logic to use informatica or any other ETL tool on top of Hadoop.

I think, effort to bring Hadoop to enterprise require diligent changes in datawarehouse reference architecture. We are going to change a lot in our Reference Architecture when we bring Hadoop into the enterprise.

Source: http://datumengineering.wordpress.com/2013/11/