# Big Data Open Source Stack vs. Traditional Stack for BI and Analytics

Part I

By Sam Poozhikala, Vice President – Customer Solutions at StratApps Inc.

4/4/2014

You may contact Sam Poozhikala at spoozhikala@stratapps.com. Follow his twitter: **@spoozhikala**. Follow StratApps: **@stratapps**

## Introduction

We get this question asked by IT and Business executives all the time. No, it is not what is big data and how can it be useful for me (that also), but more, in a simple way, can you tell me how big data compares to my traditional BI architecture stack? This is a fair question considering the dynamic nature of open source projects, marketing "Patriot Missiles" bombarding your inbox, and of course, every company touting themselves as a Big Data company. So we thought, why not share our perspective on this subject and help people understand in a way they know the traditional BI Stack and architecture.

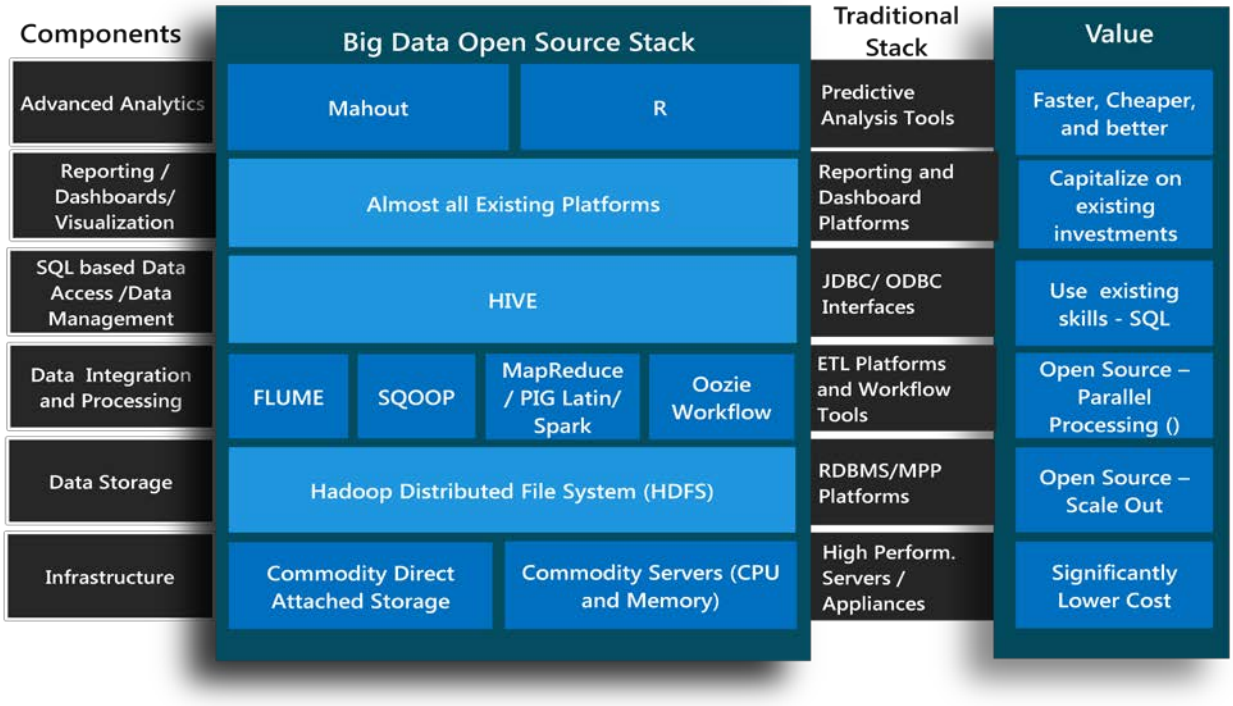
We will compare the most prominent open source initiatives in big data with the traditional stack, as well as give you a glimpse of the value proposition brought by the open source initiatives in each layer. We will also briefly explain each of the initiatives and address some of the most commonly asked questions.

In Part II, we will map each of the open source stack layers to commercial vendors and explain what they bring to the table. Our aim is that this will help you connect the dots between open source and the marketing emails you are getting flooded with!

Obviously there are further technicalities involved when you consider implementation, but that is beyond the scope as of now. If you are interested in diving deep into each component of the Big Data Open Source Stack, we have further material available at http://stratapps.com/discover.php

## Open Source Big Data Stack vs. Traditional BI/Analytics Stack

The below diagram compares the different layers of a Analytics Stack, provides you with the Open Source Initiatives that fit into the layer, and the value that you gain at each layer with the Open Source Stack.




Ok, have you spent at least 2 minutes looking at the above picture? If not, please go back and look at it again!

Let us take this from the bottom…

Infrastructure: The main difference here is that the open source big data platforms are meant to run on commodity servers and commodity storage i.e. They are all tuned to make use of the fact that you can have commodity servers and storage which are cheap, but can parallelize the processing across multiple machines of such type, thus driving increased performance at reduced cost.

Data Storage: All open source stack initiatives are file based storage. Come to think of it, even traditional RDBMS's are also files based, but you interact with the traditional RDBMS using SQL. On the other hand, you interact with files in Hadoop. What does this mean? That means, you move files into Hadoop (also known as HDFS – Hadoop Distributed File System) and you write code to read these files and write to another set of files (speaking in a simplistic manner). Yes, there is a way to manage some of this interaction using SQL "like" language. We will come to that.

Data Ingestion and Processing: Just like traditional BI, if you want to extract data from a source, all you need to do is load it and apply logic in it, the same process applies to Hadoop eco system. The primary mechanism for doing this is MapReduce, a programming framework that allows you to code your input (ingestion), logic and storing. Now, since the data sources may exist in a different format (just like traditional BI, where you may extract data from SFDC, Oracle, etc.), there are mechanism available to simplify such tasks. Sqoop allows you to move data from an RDBMS to Hadoop, while Flume allows to read streaming data (e.g. twitter feeds) and pass it to Hadoop. Just like your ETL process may be made of multiple steps orchestrated by a native or third party workflow engine, Oozie workflow provides you the capability to group multiple MapReduce jobs into functional workflow and manage dependencies and exceptions. Another alternative to MapReduce is to use Pig Latin, which provides a framework to LOAD, TRANSFORM, and DUMP /STORE data (internally this will create a set of Map Reduce code).

Hold on! There is a new top level initiative now called Apache Spark. Spark is an in-memory data processing framework that enables logic processing to be up to 100x faster than MapReduce (as if MapReduce was not fast enough!). Spark currently has three well defined functionalities, Shark (which can process SQL), Streaming, and MlLib for Machine Learning.

Data Access and Data Management: Once you have the data within Hadoop, you will need a mechanism to access it, right? That is where HIVE and SPARK come in. HIVE allows people who know SQL to interact with Hadoop using very much SQL like syntax. Under the covers, HIVE queries are converted to MapReduce (no surprise there!).

Reporting, Dashboards and Visualization: Pretty much all vendor tools can be used to create and run reports and dashboards off Hadoop. Wait, there is a caveat here. All of these tools (except those purpose built for Hadoop) use Hive and associated drivers to manage the queries. Since Hive convers the queries into MapReduce and runs it as a job, the time it takes to return the query results may be more than running the same query on a traditional RDBMS. Yes, if you have Spark, then the queries will be much faster, for sure Additionally, Hive does not support all ANSI SQL features as of now. So, you may need to consider work abounds (e.g. preprocessing some of the logic) to make sure you can get a reasonable response time.

But here is the good news. We believe the current limitations of Hive will go away very soon (maybe even in months) as the fantastic open source community continues to work on Spark to enable richer, faster SQL interaction.

Advanced Analytics: This is the area where open source beats everyone to the ground. Advanced Analytics is a combination of predictive analysis, algorithms and co-relations you need to understand patterns and drive decisions based on the suggested output. Open source initiatives like Mahout (part of Apache Projects) and R can run natively on Hadoop. What does this mean? It means predictive algorithms that may use to take days to run can now be run in minutes, and re-rerun and re-run as many times as you want (because such types of analysis required multiple iterations to reach an accuracy and probability level that is acceptable).

Now we come to addressing some of the most common questions we get!

1. Is open source big data stack ready for prime time?
   The answer is Yes. From a technology and commercial vendor support, it is very much mature to be deployed in enterprises. This is based on our own experience of implementing the stack at multiple clients and observing the stability of platform over a period of time.

2. Is this solution applicable to me?
   This is a tough question to answer, but we will make it simple – if you have less than 10TB of data for analysis/BI and/or have no major value in analyzing data from social feeds like Facebook or twitter, you should wait. But, also think if your data size is less than 10TB because you cannot keep all the data business wants due of cost? Can business drive more value if historical data were made available? And if I make that amount of data available, will it be more than 10 TB?

3. Can I basically move my Data Warehouse to Hadoop?
   We wish we could give you a black and white answer. Unfortunately, this answer will be fifty shades of grey, so to say! First of all not yet, completely (Refer to the section on Reporting, Dashboards and Visualization above). Secondly, it depends on the pain points and future vision you are addressing. There are components of this solution that can address technology pains you may be experiencing with traditional BI environment, for sure.

   However, the overall benefit of this comes when you use all layers of the stack and transform the technology and business to make use of the capabilities brought forth by this architecture.

4. Are there enough people who know this stuff?
   Unfortunately, no. There are many companies who talk about this, but there are only a few who have gotten their hands dirty. However, based on our experience and assessment, pool of people who know the open source stack is increasing. Additionally, the opex required to implement and manage Hadoop environment will be significantly lower when you add the capex savings realized realize with this solution.

5. My existing platform, storage and BI vendors claim they work on Hadoop. So can't I just go with that?
   The grey becomes even more "greyer" here. The big data hype has made many vendors to claim the compatibility. Some claims are accurate and some are not because each vendor has defined their own definition of compatibility. One simple question you can ask is " Does your offering run natively on Hadoop, i.e. does it make use of the distributed, parallel processing nodes enabled by Hadoop or Do you basically offer a way to read from/write to Hadoop"?

More Questions? Contact Sam Poozhikala at spoozhikala@stratapps.com. Follow his twitter: @spoozhikala.
Follow StratApps: @stratapps