BigDataRevealed
Investigates and Discovers Data That Positively or Negatively Impacts Your Business Performance

Lessons learned by BigDataRevealed during the Certification Process with the two top Hadoop Solutions Vendors PDFLink

How this may help you in selecting a Hadoop Solutions Vendor

My objective is to share some important findings I learned during the certification process and how they impact your company both positively and negatively. I do not proclaim to be the foremost expert on Hadoop Vendors, however I am knowledgeable on what is necessary to successfully complete a Big Data project and some of the pitfalls to avoid. I am aware of projects that have grossly exceeded budgets and timelines to discover the chosen Hadoop and third party add-on products in the end were not able to perform as needed. I suggest that you first have a clear understanding of both your business and technical needs. Then select vendors with the proper capabilities and functionality to fill those needs. I hope the information in the following paragraphs give you ideas on what is important to understand about Hadoop Vendors and also about your projects.

I've heard Big Data experts praise one Hadoop vendor over another because they offer a purer open source product, while other vendors are perceived to have a more proprietary offering with specialized add-on products. In my opinion, these should not be the first concerns you have. You should be concerned that the vendor's offerings fit your business processes and architecture, and allow you to deliver what you need within your existing processes and procedures along with any anticipated future endeavors.

Let me share a few of my observations with you. The issues below I have dealt with as a software vendor of BigDataRevealed. These are the kinds of tasks that are outside the normal data movement from legacy to Hadoop, and your basic BI and Reporting from raw data that has not yet been discovered for anomalies, incomplete data and other issues.

- Your solution may need to communicate with the HDFS File System, read and process the data within HDFS as well as read live streaming feeds of data.
- Your solution needs to consider when to process with MapReduce and or Spark.

  Much confusion swirls around the use of MapReduce and Spark. Below is a link to a Datamation article by Ken Hess on this topic:

  http://www.datamation.com/data-center/hadoop-vs.-spark-the-new-age-of-big-data.html .
- Your solution may need to process data and apply algorithms, NLP, data mining and regular expressions against this data.
- Your solution may need to write results back into HDFS with schemas defined in Hive and/or Impala and be made available for third party applications or home grown processes and for review via a GUI.
- Your solution may need to be secure. Will your third party vendor application support Kerberos, which is the growing industry standard.
- Your solution or your third party vendor may need to access data that is stored in Encrypted Zones. As an example, open source Apache Drill was unable to access this data, so BigDataRevealed has incorporated its own code and Cloudera's Impala to accomplish this task.
- Your solution may need to process millions or billions of rows. Will your third party vendor keep the data within the Hadoop Distributed File System staying within the secure Hadoop eco-system, taking full advantage of the scalability of HDFS and not exposing the data by moving it to other locations? Imagine the additional processing time that will be required if your vendor requires data be moved out of HDFS for critical data processing.

- Your solution may need to discover Personal Data, Compliance Risks, Anomalies (Outliers and Atypical data points) for all the data and NOT JUST A SAMPLE. Will your third party vendor process all entries?
- Your solution may require User definable discovery algorithms to be included. Will your third party vendor supply a mechanism to add your algorithms or will it be an addition development cost?
- Your solution may need to directly run analytics from and within major Databases (MySQL, Oracle, Teradata, DB2, etc.) enabling you to achieve Big Data processing capabilities and functionalities without having to move data into HDFS. We provide optimized MapReduce / JDBC connectors and ETL processes in Pig to help you achieve this.
    - This may be helpful in deciding what data can be moved now and what data should be reviewed or approved for movement (for sensitive data).
    - There are several Apache tools to move your data into Hadoop.

  You may desire your solution to run within existing processes. Is your third party vendor's product written modularly so that you can execute only the feature(s) you need at the

BigDataRevealed, successfully addressed each of the above processes and potential issues during the Certification process with Cloudera while staying within the framework of Hadoop and adhering to the security, encryption, methods that allow it to drill to the HDFS Record Level, with the available file system capabilities.

I summarized many of the features and capabilities of BigDataRevealed, as an example of a Hadoop large scale project and development effort, to clarify the fact that depending on the scope of your project, expectations of your deliverables, scalability needs, type and details of your graphical delivery methods, there may not be many vendors / third party vendors or possibly just one that offers everything you need to accomplish your Hadoop initiatives.

BigDataRevealed will always remain able to run on all popular Hadoop vendor platforms while being able to capitalize on Cloudera's Navigator for lineage and Impala when dealing with Encrypted Zones and data retrieval. BigDataRevealed is also completely embedded in the HDFS eco-system for maximum processing speed and efficiency.

Hadoop technology Customers need to take time with their plans and specifications and walk through each and every vendor of potential choice and run POC's for each and every aspect of each and every deliverable to assure the vendor choice (as well as third party tool choice) will work from A-Z for your project and delivery needs. I am suggesting that you start at the end of your project plan, where there is normally an assumption that all selected technologies will be meeting the project requirements. Up front due-diligence of all vendors will save time and money while improving the odds of the project implementation plan being successfully completed on time and on budget.