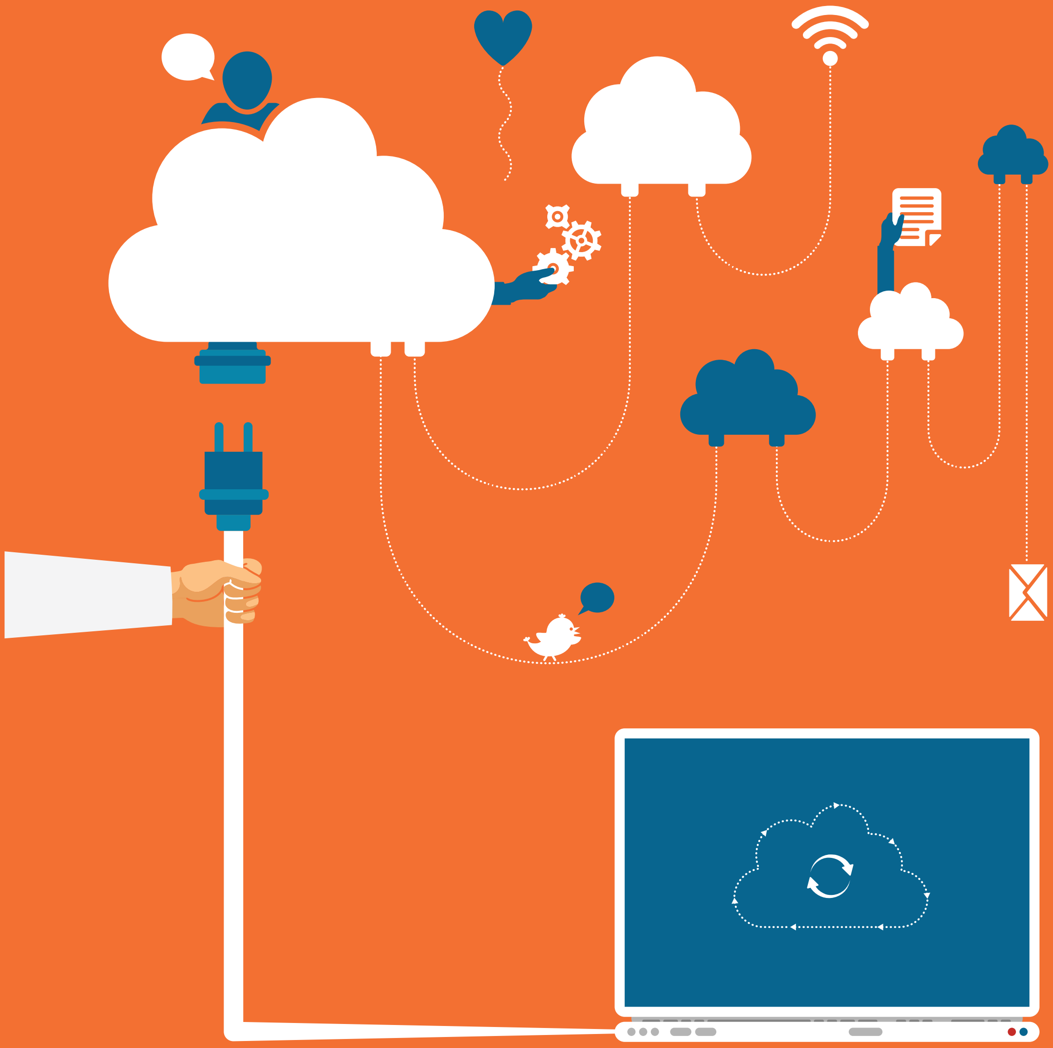
WIPRO
Applying Thought



HADOOP VENDOR DISTRIBUTIONS

THE WHY, THE WHO AND THE HOW?

Guruprasad K.N.
Enterprise Architect
Wipro BOTWORKS

Table of contents

Abstract

Hadoop has more or less become synonymous with 'Big Data' today. Hadoop is an open source project and a number of vendors have developed their own distributions, adding new functionality or improving the code base. But the many distributions also contribute to a decision complexity on which distribution to choose for your needs. Also, why have vendor distributions when there is a 'standard Hadoop distribution'"? Who are the major vendors and how do they compare? Read on to know more.

The Why - Need for Vendor Distributions

A 'standard Hadoop distribution' (Apache Hadoop) comprises:

a) MapReduce – a programming model for large scale data processing or computations in parallel

b) YARN – resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' application, providing the run time engine for MapReduce programs

c) The Hadoop Distributed File System (HDFS) – framework to manage your data

d) Hadoop Common - libraries and utilities used by the above

In addition to these, there are other solutions by Apache like Pig, Hive, Sqoop, Flume, Spark, etc., to address specific tasks or provide optimizations. Vendor distributions are useful to overcome issues, notably support related, in the standard distribution. Vendor distributions tend to focus on things such as:

1) Reliability – with prompt delivery of fixes and patches and hence are more stable

2) Support – provide technical assistance thereby enabling open source platforms for mission critical and enterprise grade usage

3) Advanced System management and Data management tools [Governance] – supplement with other tools and features mostly in the areas of security, management, workflow, provisioning and coordination. In some cases, vendors combine Hadoop with their own proprietary distributed file systems providing advantages over HDFS.

4) 'Engineered Systems' – a natural extension to the earlier point where vendors combine proprietary hardware [designed for handling big data] and software are combined to form Big Data systems.

Hadoop 1.0 had some limitations like SPOF, scalability issues beyond 4,000 nodes/tasks, ineffective utilization of resources, etc. Even before Hadoop 2.0's launch, some of these vendors already overcame such limitations in their distributions. Hadoop 2.0's YARN did bring significant architectural changes eliminating the need for these 'customized versions', but for factors mentioned above, vendor distributions still play a major role in any enterprise grade Big Data landscape.

As a side note, Hadoop 2.0 also enabled to position itself as a real time multi-use [batch, interactive, online, streaming, etc.] data application platform while Hadoop 1.0 was primarily a batch data processing solution.

The Who - Prominent Hadoop Distributions

Three of the major pure-play Hadoop distribution vendors are Cloudera, MapR and Hortonworks.

Several hardware/infrastructure vendors like Oracle, IBM, Intel, Pivotal and other companies also provide their own distributions and do their best to promote their distributions by bundling Hadoop distribution with custom developed systems referred to as 'engineered systems'. The engineered systems with bundled Hadoop distributions form the "engineered big data systems".

Reports on Big Data vendors pretty much identify these players but miss out on distinguishing between 'pure-play software' Hadoop distribution vendor v/s Engineered/Cloud-based Hadoop solutions. This distinction is very important since any service company or enterprise would first want to experiment with software only solutions and then only mature progressively on the Big Data front before investing heavily in an engineered system. When cost is a factor, software only solutions have a head-start.

The How - Comparison of Vendor Distributions

A two pronged comparison is attempted here. A high level comparison of the solution along with non-technical criteria like strategy, pedigree, customer base is attempted first thereby shortlisting a few distributions from the many in the market, followed by a detailed comparison of the components each of the distributions provide.

For high level comparison, some of the most important factors that form the basis are a) Product maturity in terms of features and compatibility with Hadoop b) Customer base c) Revenues. There are other factors like product roadmap, support, licensing and pricing, setup, management and monitoring tool support too. MapR, HortonWorks and Cloudera are almost always regarded as top three in this space. The vendor market is also changing fast with Intel announcing in Sept 2014 that it is quitting its own distribution initiative and joining forces with Cloudera.

A detailed comparison in the form of the components they bundle is provided below so that one can objectively compare the features provided.

Component	Hortonworks [HDP 2.1.5]	Cloudera [CDH 5.1.2]	MapR 4.01
File System	HDFS 2.4	HDFS 2.3.0, FUSE-DFS	MapR-FS, HttpFS
MapReduce	MapReduce 2.4	MapReduce 2.3	Propietary [MRv1, MRv2]
Non-relational DB	HBase 0.98, Accumulo 1.5.1	HBase 0.96, Apache Parquet 1.25 [columnar file]	HBase 0.98, Accumulo
Metadata	Services	HCatalog 0.5	Hive 0.12 Hive 0.12
Scripting Platform [data analysis platform]	Pig 0.12.1	Apache Pig 0.12 Apache Datafu 1.1	Apache Pig 0.12
Data Access & Query	Hive 0.13.1	Hive 0.12	Hive 0.12
Workflow Scheduler	Apache Oozie 4.0	Apache Oozie 4.0	Apache Oozie 4.0.1
Cluster Coordination	Apache Zookeeper 3.4.5	Apache Zookeeper 3.4.5	Apache Zookeeper 3.4.5
Integration with RDBMS	Apache Sqoop 1.44	Apache Sqoop 1.99, 1.44	Apache Sqoop 1.99
Integration with Streaming/Log data	Flume 1.4	Apache Flume 1.4	Apache Flume 1.5
Machine Learning	Mahout 0.9	Apache Mahout 0.8	Apache Mahout 0.9, MLLib 1.0.2

Component	Hortonworks [HDP 2.1.5]	Cloudera [CDH 5.1.2]	MapR 4.01
Hadoop UI [data integration]	Hue Talend Open Studio for Big data	Cloudera Hue 3.5	Hadoop User experience [Hue] 3.5
Cloud Services	Whirr	Apache Whirr 0.9	Apache Whirr 0.8.1
Parallel Query Execution /Interactive SQL/BI	Apache Tez 0.4.0	Cloudera Impala 1.3	Tez 0.4, Impala 1.2.3 Drill 0.5
Full Text Search		Apache Solr 4.4 Cloudera Search 1.2	LucidWorks Search 2.6.1
Non-MapReduce tasks	YARN 2.0.4	YARN 2.3.0	YARN 2.4.1
Administration	Apache Ambari 1.5.1	Cloudera manager	MapR Control System [MCS] – Heatmap, Job Metrics,
Installation	Apache Ambari 1.5.1	Cloudera manager	
Monitoring	Ganglia 3.5.7, Nagios 3.5.0		
Authorization		Apache Sentry 1.2	
Core Hadoop			2.4.1
Streaming			Spark Streaming 1.0.2
Others	Storm 0.9.1 Apache Falcon 0.5.0 [data management platform] Apache Knox 0.4.0 [REST API for Hadoop clusters] Apache Phoenix 4.0.0 [SQL Skin over HBase – JDBC driver]		Spark Streaming 1.0.2
Editions		Cloudera Express Cloudera Enterprise	Standard [M3], Enterprise [M5], Enterprise DB edition [M7]

Observations

- The sheer number of components and solutions bundled in the vendor distributions is justification enough [and an important reason] for enterprises to choose a vendor distribution over standard distribution. Imagine the "version-hell" [working out which versions of one component works well with versions of another component, etc.] an enterprise has to go through if they were to do this on their own. The vendor distributors do a great job of testing and ensuring compatibility amongst different component versions and also keeping them up-to date by regular releases. This is not to say that this should be the sole reason to choose a vendor distribution. Of course, there are other factors and features which they bundle which make it enterprise ready
- Hortonworks bundles mostly open-source components and there is nothing proprietary in its stack. Cloudera's management and administration features are proprietary as also Impala [its improvement over Hive] while MapR's core [file system] itself is proprietary
- From a market perspective, Cloudera clocked ~$70MM, while MapR and Hortonworks clocked $30MM and $25MM respectively last year. These 3 are still fighting for Hadoop supremacy in terms of marquee customers, funding and market share. These are working out alliances too with Cloudera being bundled by Oracle in its Big Data Appliance and RedHat preferring Hortonworks. MapR with its proprietary stuff seems to be falling behind in terms of popularity as well as usage, but is still a major player to contend with. At present, Cloudera has a larger market share followed by MapR and Hortonworks
- Cloudera uses core Hadoop and open source platforms related to Big Data but innovates where it is needed – Administration/Management and efficient execution engines. It built Impala, a faster SQL Engine for Hadoop, but with Apache Spark filling that gap it might have to rethink if it wants to pursue Impala at all. Initial tests have shown Spark performing better than Impala. It has 200+ customers and is the market leader in all aspects [The cells marked in yellow highlight the proprietary components/products used in the bundle]
- Hortonworks is completely reliant on open-source for all its needs and is betting big on Ambari for administration/management functionality too
- MapR scores heavily in Analyst firm ratings and is second to Cloudera in terms of market penetration. Its only [perceived?] liability is its proprietary file system and MapReduce model which makes it difficult for customers to have an option of exiting the platform at a later stage. This is also making it difficult to adopt to changes in the core Hadoop platform and was the last to shift to Hadoop v2.0
- As can be seen from the table, in terms of features and components, there is very little to differentiate each of the vendors
- Other vendors are also building up their capabilities and catching up with these 3 – notably Pivotal Greenplum, Microsoft HDInsight, Teradata's SQL-H and IBM's Infosphere BigInsights

Future Trends Influencing Vendor Distributions

- Integration with data analysis, BI and reporting solutions – Integration with packages like Tableau, Spotfire, etc. will pick up so that a comprehensive tool set is provided to customers. Right now, distributions just package the components that one has to use to build their own algorithms and provide means to visualize the results. Integration with BI and reporting solutions will fill that gap. Integration with self-service BI tools like Apache Drill will also pick up
- More focus on performance improvements – Earlier the focus was on getting all of the data into one place and making some sense out of it. With Hadoop v2.0, it is no more batch processing apps only that are using Big Data. Interactive apps demand speed and hence there would be more innovations to satisfy the need for speed. Spark, Shark, etc. are initiatives in that direction from the open source community. Expect the vendors to do better
- Need for security – With more interactive apps and self service BI, etc., coming into play, security would now be given more prominence than when only batch apps were run

Conclusion

The only difference between Cloudera and Hortonworks is in its management/administration component – Cloudera manager v/s Apache Ambari. Of course, Hortonworks has only the free model, while Cloudera provides both free and premium models. Cloudera does have a large customer base and market share. Making either of these two choices are OK and it depends on the scale of adoption today. It would not be very difficult to move from one distribution to another either.

About the Author

Guruprasad K.N., Enterprise Architect, Wipro BOTWORKS

Guruprasad K. N. works as an Enterprise Architect in the BOTWORKS division of Wipro and has a cumulative experience of 19 years in IT industry. Apart from architecting systems, he is very passionate about developing solutions and accelerators in the area of Big Data, Internet of Things [IOT], Machine Learning and other emerging technologies. He has varied experience in the telecom, healthcare, online payment, eCommerce, and energy domain. Guru has 4 patents in the areas of healthcare and IT transformation and has authored papers in the ACM journal.

About Wipro Ltd.

Wipro Ltd. (NYSE:WIT) is a leading Information Technology, Consulting and Business Process Services company that delivers solutions to enable its clients do business better. Wipro delivers winning business outcomes through its deep industry experience and a 360 degree view of "Business through Technology" - helping clients create successful and adaptive businesses. A company recognized globally for its comprehensive portfolio of services, a practitioner's approach to delivering innovation, and an organization wide commitment to sustainability, Wipro has a workforce of over 150,000, serving clients in 175+ cities across 6 continents.

For more information, please visit www.wipro.com

WIPRO
Applying Thought

DO BUSINESS BETTER

WWW.WIPRO.COM

CONSULTING | SYSTEM INTEGRATION | BUSINESS PROCESS SERVICES

WIPRO LIMITED, DODDAKANNELLI, SARJAPUR ROAD, BANGALORE - 560 035, INDIA TEL : +91 (80) 2844 0011, FAX : +91 (80) 2844 0256, email : info@wipro.com

North America Canada Brazil Mexico Argentina United Kingdom Germany France Switzerland Nordic Region Poland Austria Benelux Portugal Romania Africa Middle East India China Japan Philippines Singapore Malaysia South Korea Australia New Zealand

IND/B&T/JAN - JUN 2015