Compare and contrast the 3 predominant Big Data platform vendors in the marketplace?

Hadoop is an open source project to handle bigdata. Different features can be personalized for different users, and present them in the Apache repository because it is an open source project. Cloudera, Horton, and MapR became major players on top of Hadoop framework to offer enterprise-ready Hadoop distribution.

Comparison

	Cloudera	Horton	MapR
Distribution	• Compatible with Apache Hadoop • Provide to ensure security and stability • Provide paid training, consulting, technical services • Established communities that actively participate and help with the problems • Support the network file system protocol which is possible to mount onto another server.		
Name	• CDH (lower capability version) • Cloudera Altus Director	• HDP(platform) - based on Apache Hadoop, Hive, Spark • HDF(flow) - Apache NiFi, Storm, Kafka	• MapR Converged Data Platform
Management	• Rich User friendly interface • Easy to administer Hadoop • Proprietary management software • Cloudera Search for real-time access of products, and Impala, an SQL query handling interface	• Easy to install and free • First Hadoop Distribution that supports Windows platform. • The Ambari Management interface on HDP • Service only distribution • Hive became faster through its new Stinger project. • Avoids vendor lock-in • Focus on enhancing the usability • Apache Solr for searches of data	• Fastest Hadoop distribution with multi node direct access. • The Event Store for Apache Kafka allows to run heavy streaming workload on one cluster in production.

	Cloudera	Horton	MapR
License	• Possesses a commercial license	• holds an open source license • Licensing cost and Horton Data Cloud at AWS is higher when compared to other distribution partners	• M3 – free • M5 – degraded availability feature • M7 - purpose-built rewrite of HBase that implements the HBase API directly in the file-system layer
Cost	• Paid service with free trial for 60 days • Software license fee and cloud service fee are separate.	• Free to use • Can add paying subscription for troubleshooting, training, upgrades, diagnosis	• Free version – Community Edition, community forum support, unlimited production use including Hadoop, Spark • Converged Enterprise Edition (MapR EE) – extensive deployment, disaster recovery, premium support, customized selection of features.
Software	Sells commercial software on top of open source Hadoop distribution - Hybrid	100% open source and offers only Apache foundation certified software	Support Apache open source projects, but also focused on develop proprietary software to improve product reliability
Business Strategy	Traditional software provider that profits from product sales and enhanced technology based on open source framework.	Embed Hadoop into existing data platforms and solely rely on open source development	Niche market to organizations with demanding production needs
Window support	Not a native component but can be run on windows server. Cloudera Quickstart VM required	Included as a native component on the windows server	Window support in a VM environment

Features

	Cloudera, Horton	MapR
Security & Governance	- Used to not have expressive authorization, catching up with unified authentication - Cloudera Shared Data Exchange (SDX) to improve analytics, security and governance	- Platform-level security - More advanced in security and governance - Enterprise Data Catalog across the enterprise data platform
Cloud Management	Harder to synchronize data across cloud and edge deployment	Global namespace handles to synchronize all data and a single view into all data
Recovery	Disaster recovery must be done manually	Built in multi-tenancy, disaster recovery
File System	- Master-slave architecture - HDFS component with DataNode and NameNode architecture - HDFS is written in Java, and need to run in the JVM. In order to write to an append-only file system, we need to write to temporary space on the Linux file system, which supports both read/write capabilities then can send it back to the gateway which writes it into HDFS. This layer problem has been addressed and on progress of improvement	- Instead of HDFS, proprietary solution MapRFS used to provide more efficient management of data and reliability. - Does not rely on Linux file system - You can run any application that has file system requirements on MapRFS - MapR has accesses to raw disk drives. - The MapR filesystem is very easy to integrate with other Linux filesystems
Data Science and AI	Access to Python ML libraries requires a separate cluster, resulting in data copies which can cause security and lineage issues – This has been improved in 2017 for data scientist in a self-service style.	Support for open APIs like POSIX, allow AI and ML to run on same cluster as your analytics

	Cloudera	Horton	MapR
Major Open Source Project	• Hadoop • Accumulo • Flume • HBase • Hive • Impala • Kafka • Pig • Sentry • Spark • Sqoop • HUE	• Accumulo • Ambari • Atlas • Falcon • Flume • Hadoop • Hadoop HDFS • Hadoop MapReduce • YARN • HBase • Hive • Kafka • Knox Gateway • Metron • Nifi • Oozie • Phoenix • Pig • Ranger • Slider • Storm • Solr • Spark • Sqoop • Storm • Tez • Zeppelin • Zookeeper • Druid	• Hadoop • Hbase • Pig • Hive • Mahout • Spark • Spark Streaming • Drill • Solr • Zookeeper • Sqoop • Oozie • HUE
Special Technologies	• Cloudera Search • Cloudera Enterprise Data Hub • SDX – Cloudera shared data exchange • Cloudera Operational DB • Cloudera Work Load XM		• HttpFS • MapRFS • MapR POSIX client • MapRDB • MapR Streams • MapR Event Store • MapR XD – MapR Distributed File and Object Store Cloud-Scale Data Store

Partnership

Cloudera	Horton	MapR
Cloudera maintains a partner ecosystem that includes over 2,300 firms; they fall into the categories listed below • Analytics & Business Intelligence: MicroStrategy, Oracle, SAS • Cloud Partners: Amazon Web Services, Google Cloud Platform, Microsoft Azure • Data Integration Partners: Attunity, Cognizant, SAP • Hardware Vendors: Dell, EMC, Fujitsu • Managed Service Providers: CenturyLink, CSC, Teradata • Resellers: Cisco, NetApp, Hewlett-Packard Enterprise • Software Vendors: Datameer, Digital Reasoning Systems • System Integrators: Accenture, Capgemini, Deloitte	There are over 1,600 partners • Technology Partner (ISV/IHV) Program: independent OEMs, hardware, and software solution providers to develop, test, deploy, and support joint products with Hortonworks offerings - Hewlett-Packard Enterprise and EMC2. • System Integrator/Consultant Program: system integrators and consultants who provide consulting, design, and/or deployment services to optimize their services through use of HDP - Accenture, Capgemini, and EY. • Strategic Reseller Program: agreement with resellers for the sale of Hortonworks subscriptions and other services to their clients - World Wide Technology and Immix Group. • Training Delivery Partner Program - provide training for high-tech software firms, in which they receive training themselves on the use of HDP.	Partnership to generate a broad variety of solutions • Systems Integrators/Consultants: Required to have an active big data practice to join • Authorized Resellers: an in-person selling model, and complementary products and services. • OEM Partners: manufacturers to integrate its platform. • Technology licensing agreement with EMC • Partner with AWS for upgraded version of Amason's Elastic MapReduce(EMR) service for extra cost • Google's technology partner. MapR broke the speed record on Google's Compute Engine. • First Certified Training in EMEA • Support on the Cisco Unified Computing System platform

Pros

Cloudera	Horton	MapR
• 1st enterprise Hadoop distributor and biggest market player • Automate the installation process • Reduced deployment time • Display real time nodes count • Able to add new services to a running Hadoop cluster • Multi cluster management • Provide Node templates with varying configuration • Faster updates and bug fixes to the products as they have Apache committers • Easy to install on VM and well documented with examples • Cloudera Impala is fast and free to query the HDFS, but not free to do that at Amazon EC2 and Windows Azure.	• Mature integration in various applications. • Flexibility in case of new migration • Easy for beginners to start with Hadoop. • Pure open source platform in Hadoop ecosystem • More choice of open source tools. • The convenience of Ambari UI and API for building, deploying and managing the cluster makes it relatively easy to get started. • With YARN and Spark, you can mix different nodes for storage and compute and master nodes to manage loads. • Horton sandbox can be installed for learning and development purposes • Increasing market share as a Hadoop distributor	• MapR is fast, reliable, scalable, dependable, and higher throughput. • Improved reliability of the system such as bug fixing. • Expand functionalities • More production ready • Quickly respond to market needs • Complete data protection • Solid central configuration • Real time streaming data management • Built in connectors to existing systems • Enterprise quality engineering • Largest deployment in the financial services industry, • Easy integration with other Linux file system - No need to set up NFS gateways to use command line utilities native to Linux. • Abundant free learning materials

Cons

Cloudera	Horton	MapR
• Slower than MapR Hadoop Distribution for more critical or dynamic workloads, • Cloudera technical support is acceptable, though not overly customer oriented. • Still needs some improvements on central configuration • Cloudera still have full control even though it's based on Hadoop because some components are privately owned • All is Linux based so it's better for users to know Linux commands • GUI is missing • Matching Cloudera product version with CentOS can be tricky	• It is free to install and get deployed. But it's not free to get support. You have to reply on the community and your own research for problems. • Functionality is behind compared to Cloudera and MapR • Version upgrades have been challenging. • Compatibility issues need to be resolved manually. • It is not ideal to process streaming data. And you need to configure data process yourself. • VM setup can be more difficult than Cloudera, and it is very slow to process data in VM environment. • Ambari management interface is just basic, and doesn't have rich features compared to Cloudera	• Since it's based on proprietary technologies, it would be harder to change if we want to switch to a different distribution later. • Poor name recognition in Hadoop industry compared to Horton and Cloudera • Deployment mechanisms are more difficult to use especially when it can be automated. • Documentation could be improved. • The MapR web UI console is pretty basic. It can make it easier to administer and manager clusters • MapRFS can be improved to work better with HBase. Hbase writing throughput is tremendously high that it can actually slow down to improve overall system throughput on MapRFS

New Focuses and use cases

As we studied earlier, all 3 distributors have had weaknesses. The trend is that they try to improve on their short comings but the strengths of their competitors, and thrive to be an industry leader.

Cloudera	Horton	MapR
The next focuses for Cloudera would be machine learning, analytics, and clouding which have not been their forte historically. Their main customers have been more about adding big data ingestion from new data sources to their Hadoop environment on top of clients' existing RDBMS. 20% of its customers are already using on cloud, and it is expected to grow more within a few years. Cloudera released Altus Data Science which will bring machine learning workload by R and Python at platform environment instead of infrastructure. This has been their main weakness compared to MapR as the workload had to go around with Linux. SDX has been integral part of security, governance and management. But it's has been only available for on-premise customers. But now it made it capable of managing across different clusters.	Horton also released new HDP with more extensive support on functionality and performance. It allows for more use cases for performance increase for SQL interactive queries, and optimizing existing data warehouse without requiring all data to be reloaded. Also BI service level and data management has been enhanced with Hive and Spark. Similar to Cloudera, they want to make sure to have bigger pie in Cloud, streaming analytics, and processing big data in motion which have been their weaknesses at the expense of keeping them all in open source projects. They are shifting their image from Hadoop distributor, to their new streaming data platform called HDF and data architecture company. Their major clients already adopted HDF and their new growth is around streaming data. They also developed streaming analytic applications to close the gap with business users and make it simple to set up their HDF. with cloud AWS and Azure and integrate them into cloud data stores. We can tell their focus is not just data ingestion and processing, but also data utilization in motion.	On top of speedy performance which has been their main competitive advantage, they also put their energy on cloud storage, application portability, analytical and streaming capabilities and machine learning that requires new architectural innovations. They also enhance even more on data platform security with one click installation. Recently Cloudera and Horton got merged and their marriage seems natural as their capabilities are complementary each other and both distributors rely on more open source technologies compared to MapR. Plus, Cloudera and Horton caught up in many areas which only MapR could do in the earlier age of Hadoop. It would be interesting to find out how MapR could breakthrough between 3 big brothers in Hadoop industry.