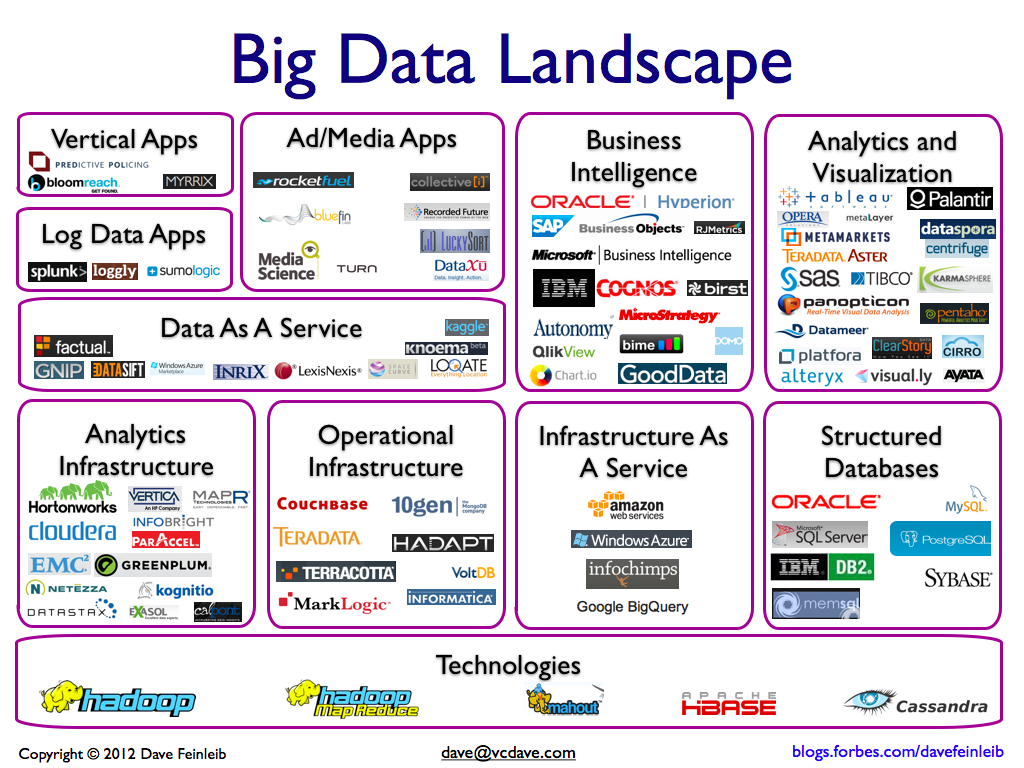
# Top BigData Platforms



# Các sản phẩm là thành phần của Big Data Ecosystem



# Tiêu chí đánh giá

1010data puts analytics in the cloud 
 
Analytical DBMS: 1010data columnar analytical database. 
In-memory DBMS: None. 
Stream-analysis option: None. 
Hadoop distribution: None.  
Hardware/software systems: Not applicable.
New York-based 1010data launched its analytical, private-cloud service way back in 2000, building a base of customers on Wall Street. Marquis customers include NYSE Euronext and a number of big banks, but the company has also branched out into retail, CPG, gaming, healthcare, government, and telecommunications.
1010data's columnar database supports massively parallel processing for scalability, but it's a proprietary design with its own query language that supports a subset of SQL functions plus broader query types including graph and time-series analyses. It also handles semi-structured data such as social network and machine data. Beyond the database, the company offers a complete stack including data integration, reporting, and data-visualization tools, as well as advanced analytic functions including statistical analysis, optimization, and machine learning.
1010data's private-cloud approach relieves customers of the burden of managing and scaling infrastructure. Centralized management and access controls and APIs support integration with back-end systems as well as broad access to information with 'HIPAA-grade' security. The company has more than 250 customers. In contrast to a cloud provider such as Amazon, which delivers standardized (very-low-cost) services to tens of thousands of customers, 1010data is a custom services provider that crafts private-cloud applications and capabilities matched to customer needs.

# Comparation

## 1. Gardner

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Region | Detail | MapR | Cloudera | HortonWorks | HP Vertica | Oracle Big Data |
| Summary |  | Đánh giá cao ở môi trường Analytics | Đánh giá cao ở tính mở, tốc độ, khả năng tích hợp | Đánh giá cao ở khả năng hỗ trợ nhiều môi trường phát triển | Đánh giá cao khi tích hợp với nền tảng cơ bản của BigData | Đánh giá cao ở môi trường Enterprise, trước đó doanh nghiệp đã sử dụng các sản phẩm khác của Oracle |
| Open |  |  |  |  |  |  |
|  | *All (granular) data* | Normal; Thay thế HDFS bằng NFS | Normal; Tương tự Apache Hadoop thuần | Good; Hỗ trợ đa nền tảng, vẫn hỗ trợ đầy đủ Apache Hadoop Ecosystem | Good; Rất nhiều Partner tham gia quá trình phát triển Vertica, được đánh giá cao ở mảng Data Integration | Good; Hỗ trợ đa dạng các nguồn dữ liệu từ nhiều môi trường: Mobi, network, local |
|  | *Off-road analysis* |  |  |  |  |  |
|  | *Enterprise integration* | Normal; Tương tự Apache Hadoop thuần | Good; Nếu sử dụng bản Cloudera Enterprise với Cloudera Enterprise Management | Good; Mục tiêu của hệ thống hướng tới dòng Enterprise | Good; Bổ sung tốt cho nền tằng Apache Hadoop thuần, MapR hoặc Cloudera | Good; Tích hợp tốt với toàn bộ các sản phẩm của Oracle: Oracle Database and MySQL; Oracle Sun; Oracle Event Processing, Oracle Coherence, Oracle NoSQL, GoldenGate and Data Integrator (data integration), Oracle Business Analytics, and Oracle Real-Time Decisions, ... |
| Easy |  |  |  |  |  |  |
|  | *Visual and interactive* | Normal; Công cụ quản lý sơ sài | Nomal; Có giao diện quản lý trực quan | Normal; Ứng dụng quản lý tương đối trực quan, chuyên nghiệp với Apache Ambari | Good; Sử dụng HP Vertica Management | Good; Oracle Enterprise Management |
|  | *No Programming or DB knowledge needed* | Good; Phù hợp để lựa chọn cho các dự án recommendation engines, fraud-detection, and predictive applications | Normal; Tương tự Apache Hadoop thuần | Weak; Chỉ hỗ trợ nhiều môi trường phát triển, không hỗ trợ tới mức Tool tương tác người dùng | Good; HP Vertica Designer | Good; Nhưng chỉ phục vụ cho Nền tảng Oracle Business Analytics |
|  | *Fully managed service* | Normal; Tương tự Apache Hadoop thuần | Good; Sử dụng Cloudera Service Manager | Near Good; Sử dụng Apache Ambari | Good; HP Vertica Management | Normal; Không tích hợp sâu với Hadoop Ecosystem |
| Fast |  |  |  |  |  |  |
|  | *Fastest query speeds* | Good; Theo đánh giá là nhanh hơn HDFS vì xây dựng trên NFS | Good; Với Apache Impala | Good; Đang phát triển Stinger | Good; High RAM-to-disk ratios | Good; Oracle Timesten, ... |
|  | *Almost immediate delivery* | Weak; Chưa được đánh giá cao ở môi trường Enterprise | Good; Sử dụng Cloudera Service Manager | Good; Hỗ trợ tốt môi trường Cluster nhưng phải trả phí | Good; HP Vertica Management | Normal; Không tích hợp sâu với Hadoop Ecosystem |
| Sophisticated Analysis |  |  |  |  |  |  |
|  | *Advanced statistics and modeling* | Good; Nền tảng tốt cho BI | Normal; Tương tự Apache Hadoop thuần | Normal; Vẫn là hệ thống hướng đến môi trường chung, nhưng hỗ trợ nhiều môi trường phát triển | Weak; HP Vertica không được đánh giá cao so với các sản phẩm tương tự của: IBM, SAP & Oracle | Good; Thông qua Oracle Business Analytics |
|  | *Unrestricted joins on disparate data* | Normal; Tương tự Apache Hadoop thuần | Weak; Không hỗ trợ Join tương tự Apache Hadoop thuần | Good; Với Stinger | Weak; | Good; Là thế mạnh của Oracle Database hiện tại |
|  | *No need for sampling or aggregation* | Good; Nền tảng tốt cho các hệ thống Dự đoán, phân tích dữ liệu | Normal; Tương tự Apache Hadoop thuần | Weak; Vẫn dựa trên Apache Hadoop | Weak | Good; Oracle Business Analytics |
| Of Big Data |  |  |  |  |  |  |
|  | *Trillions of records* | Good; Thay thế HDFS bằng NFS | Normal; Tương tự Apache Hadoop thuần | Good; Phát triển Apache Tez | Normal | Good; |
|  | *Structured or semi-structured* | Normal; Tương tự Apache Hadoop thuần | Normal; Tương tự Apache Hadoop thuần | Good; Tương tự Apache Hadoop | Normal | Good; |

## 2. Technical Comparation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Region** | **ApacheHadoop** | | **MapR** | **Cloudera** | **IBM BigInsights** | **HortonWorks** |
| **CORE HADOOP** |  |  | |  |  |  |
|  | Core Hadoop, YARN | Core Hadoop, YARN | | Core Hadoop, YARN | Core Hadoop, YARN | Core Hadoop, YARN |
| **BATCH** |  |  | |  |  |  |
|  | MapReduce | MapReduce | | MapReduce | IBM Advanced MR (IBM Symphony) | MapReduce |
|  | Hive | Hive | | Hive | Hive | Hive |
|  |  | Tez | |  |  | Tez |
|  |  | Storm on YARN | |  |  | Storm on YARN |
|  | Pig | Pig | | Pig | Pig | Pig |
|  |  | Cascading | |  |  | Cascading |
|  |  | Spark | | Spark |  | Spark |
|  |  |  | |  | Lucene |  |
|  |  |  | |  | HCatalog |  |
| **INTERACTIVE SQL** |  |  | |  |  |  |
|  |  | Impala | | Impala | Jaql |  |
|  |  | Drill | |  |  | Drill |
| **NOSQL & SEARCH** |  |  | |  |  |  |
|  | HBase | HBase | | HBase | HBase + ColumnStore | HBase |
|  |  | AsyncHBase | |  |  |  |
|  |  | Search | | Solr |  | Solr |
| **GRAPH** |  |  | |  |  |  |
|  |  | GraphX | |  |  |  |
| **MACHINE LEARNING** |  |  | |  |  |  |
|  | Mahout | Mahout | | Mahout |  | Mahout |
|  |  | MLlib | |  |  |  |
| **ANALYTIC ENGINE** |  |  | |  |  |  |
|  |  |  | |  | R |  |
|  |  |  | |  | Adaptive Algorithms |  |
|  |  |  | |  | Text Processing Engine |  |
|  |  |  | |  | Extractor Library (AQL + HIL) |  |
| **STREAMING** |  |  | |  |  |  |
|  |  | Spark Streaming | |  | Spark Streaming |  |
| **DATA TOOLS** |  |  | |  |  |  |
|  |  | HttpFS | |  |  |  |
|  | Sqoop | Sqoop | | Sqoop | Sqoop | Sqoop |
|  | Flume | Flume | | Flume | Flume | Flume |
| **COORDINATION** |  |  | |  |  |  |
|  | Oozie | Oozie | | Oozie | Oozie | Oozie |
|  | ZooKeeper | ZooKeeper | | ZooKeeper | ZooKeeper | ZooKeeper |
| **GUI, CONFIGURATION, MONITORING** |  |  | |  |  |  |
|  | Ambari | Management (MapR Control System (MCS)) | | Cloudera Management | InfoSphere BigInsights Mng | Ambari |
|  | Hue | Hue | | Hue |  | Hue |
|  |  | Whirr | |  |  | Whirr |
|  |  |  | |  | Apps |  |
|  |  |  | |  | Text Analytics |  |
|  |  |  | |  | BigSQL |  |
|  |  |  | |  | Workflow |  |
|  |  |  | |  | Pig & Jaql |  |
| **VISUALIZATION & DISCOVERY** |  |  | |  |  |  |
|  |  |  | |  | BigSheets |  |
|  |  |  | |  | Dashboard & Virtualization |  |
| **ADMINISTRATION** |  |  | |  |  |  |
|  |  |  | |  | Admin Console |  |
|  |  |  | |  | Monitoring |  |
| **FILE SYSTEM** |  |  | |  |  |  |
|  | HDFS | NFS (Distributed Storage) | | HDFS | HDFS + GPFS-FPO | HDFS |
| **INTEGRATION** |  |  | |  |  |  |
|  |  |  | |  | JDBC |  |
|  |  |  | |  | Netazza |  |
|  |  |  | |  | DB2 |  |
|  |  |  | |  | Streams |  |
|  |  |  | |  | DataStage |  |
|  |  |  | |  | Guardium |  |
|  |  |  | |  | Platform Computing |  |
|  |  |  | |  | Cognos |  |
|  |  |  | |  |  | WebHDFS |
|  |  |  | |  |  | Talend, Camel |

# Details

## 1. MapR sprints ahead on performance

* + **Analytical DBMS: HBase; supports Drill, Hive, Impala, Shark, and other (non-DBMS) SQL-on-Hadoop options.**
  + **In-memory DBMS: MapR touts in-memory performance through (nib-DBMS) open-source projects Drill and Shark.**
  + **Hadoop distributions: MapR M3, MapR M5, MapR M7.**
  + **Stream-processing technology: MapR supports streaming analysis through Storm and through an integration with Informatica HParser.**
  + **Hardware/software systems: Hardware configurations available through partners including Cisco, HP, IBM, and NetApp.**

MapR sprints ahead on performance 
 
Analytical DBMS: HBase; supports Drill, Hive, Impala, Shark, and other (non-DBMS) SQL-on-Hadoop options. 
In-memory DBMS: MapR touts in-memory performance through (nib-DBMS) open-source projects Drill and Shark. 
Hadoop distributions: MapR M3, MapR M5, MapR M7.  
Stream-processing technology: MapR supports streaming analysis through Storm and through an integration with Informatica HParser. 
Hardware/software systems: Hardware configurations available through partners including Cisco, HP, IBM, and NetApp.
MapR marches to the beat of its own drum, replacing bits and pieces of the Hadoop framework to deliver higher performance or to fill gaps in functionality. Early on, it replaced HDFS with an alternative based on the Network File System (NFS) to ensure high availability. In a tie between NFS and Informatica HParser software introduced in 2012, MapR introduced an option for stream processing on top of Hadoop. The 2013 MapR M7 Hadoop distribution addresses weakness in HBase by doing away with region servers, table splits and merges, and data-compaction steps. MapR also implemented its own architecture for snapshotting, high availability, and system recovery.
With M7, MapR also introduced optional LucidWorks Search software on top of Hadoop for building out recommendation engines, fraud-detection, and predictive applications. MapR promotes Apache Drill as its SQL-on-Hadoop option of choice, but it pragmatically touts open-source and commercial alternatives including Apache Hive, Impala, Shark-on-Spark, Hadapt and others, perhaps responding to rivals who slam MapR's go-it-alone ways.
The community addresses Hadoop's squeakiest wheels at its own pace, but MapR seems to thrive on moving ahead with commercial alternatives with the promise of better performance.

## 2. Cloudera Platform

* + **Analytical DBMS: HBase, and although not a DBMS, Cloudera Impala supports SQL querying on top of Hadoop.**
  + **In-memory DBMS: Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop.**
  + **Hadoop distributions: CDH open-source distribution, Cloudera Standard, Cloudera Enterprise.**
  + **Stream-processing technology: Open-source stream-processing options on Hadoop include Storm.**
  + **Hardware/software systems: Partner appliances, preconfigured hardware, or both available from Cisco, Dell, HP, IBM, NetApp, and Oracle.**



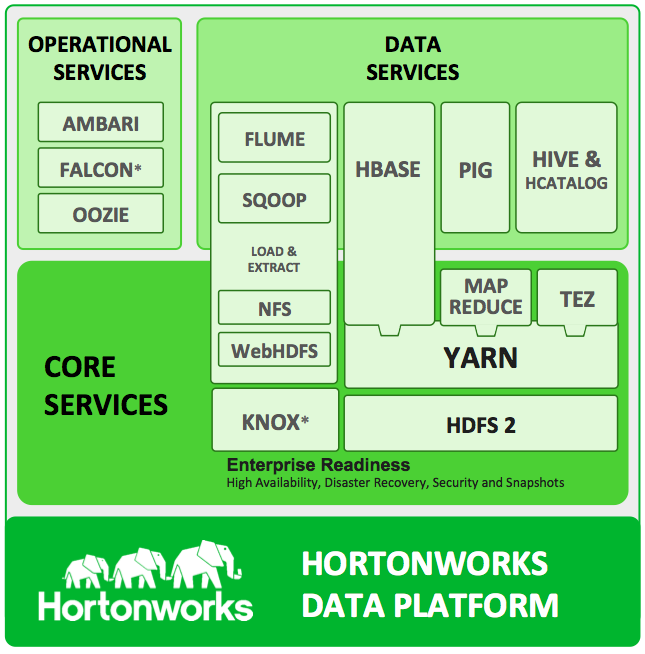
Cloudera’s revolutionary data management platform is designed specifically to address the opportunities and challenges of Big Data. Cloudera combines Apache Hadoop with a number of other open source projects to create a single, massively scalable system where you can unite storage with an array of powerful processing and analytic frameworks. By uniting flexible storage and processing under a single management framework and set of system resources, Cloudera delivers the versatility and agility required for modern data management – where you can ingest, store, process, explore and analyze data of any type or quantity without migrating it between multiple specialized systems.

* Unified Data – run transformations, analysis and applications against a common data set
* Unified Metadata – all jobs and queries leverage a single metadata model
* Unified System – all data operations run on a single set of nodes
* Unified Security – a single model for security, data access and compliance
* Unified Management – centralized system and data management

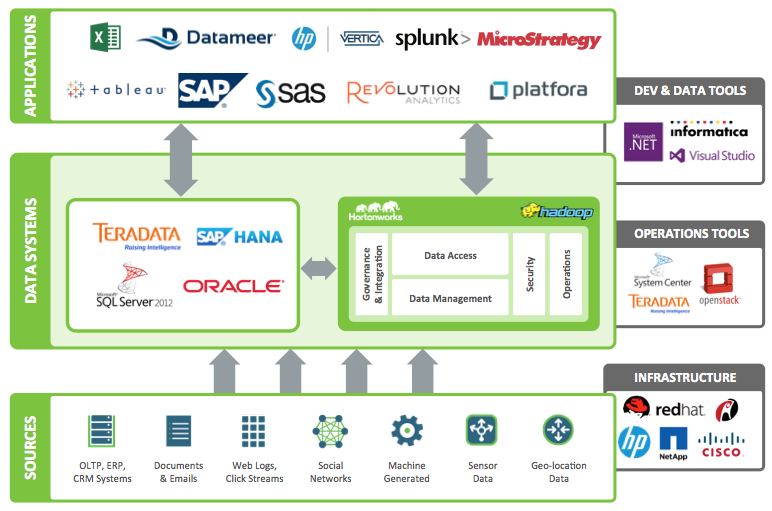
## 3. Horton Works

* + **Info**
    - **Analytical DBMS: HBase; although not a DBMS, Hive is Hortonworks' option for SQL querying on top of Hadoop.**
    - **In-memory DBMS: Although not a DBMS, Apache Spark supports in-memory analysis on top of Hadoop.**
    - **Hadoop distributions: Hortonworks Data Platform (HDP) 2.0, HDP for Windows, Hortonworks Sandbox (free, single-node desktop software offering Hadoop tutorials).**
    - **Stream-processing technology: Open-source stream-processing options on Hadoop include Storm.**
    - **Hardware/software systems: Partner appliances, preconfigured hardware, or both available from HP, Teradata and others.**

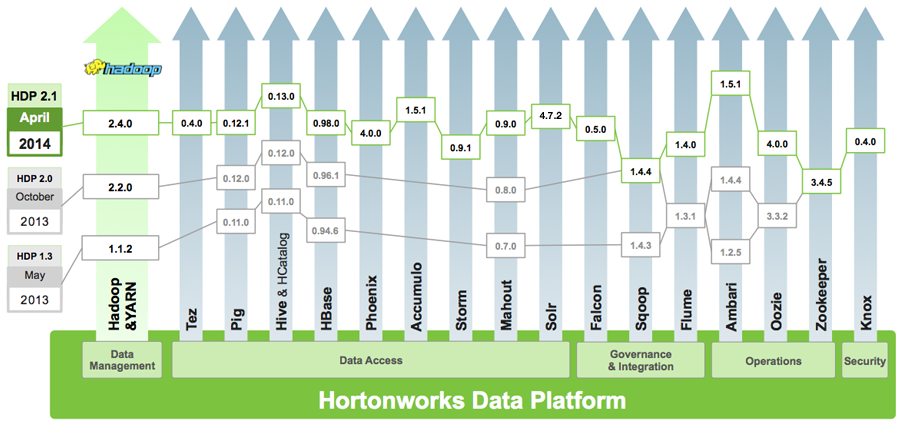
* Structure



* Ecosystem



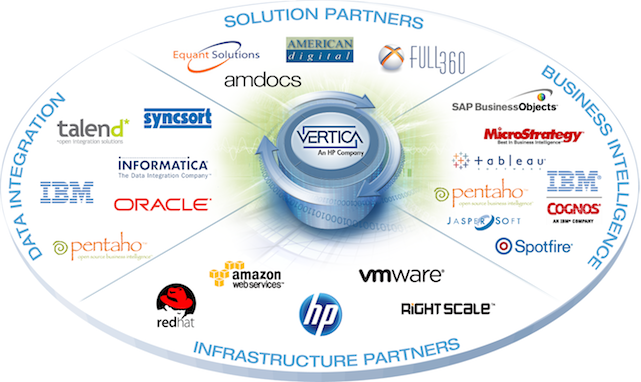
* Milestones



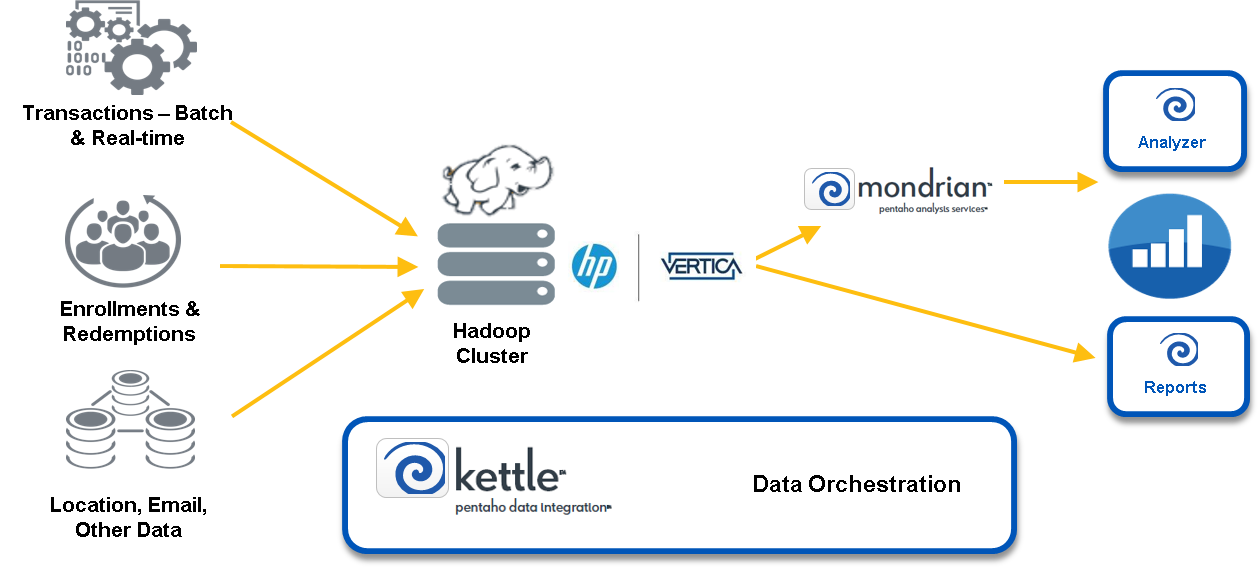
## 4. HP Vertica

* + **Info**
    - **Analytical DBMS: HP Vertica Analytics Platform Version 7 (Crane release).**
    - **In-memory DBMS: Vertica is not an in-memory database, but with high RAM-to-disk ratios the company says it can ensure near-real-time query performance.**
    - **Hadoop distribution: None.**
    - **Stream-processing technology: None.**
    - **Hardware/software systems: HP ConvergedSystem 300 for Vertica, plus a choice of reference architectures for Cloudera, Hortonworks, and MapR Hadoop distributions.**

* Partner



* **Structure**



## 5. Oracle Big Data

* + **Info**
    - **Analytical DBMSs: Oracle Database, Oracle MySQL, Oracle Essbase.**
    - **In-memory DBMS: Oracle TimesTen, Oracle Database 12c In-Memory Option (announced in 2013 without details, roadmaps, or release dates).**
    - **Stream-analysis option: Oracle Event Processing.**
    - **Hadoop distribution: Resells and supports Cloudera Enterprise.**
    - **Hardware/software systems: Exadata, Exalytics, Oracle Big Data Appliance.**
  + **Structure**

