Expansion of Data Warehouse System using Addition of Big data

Darshil Gohel

University of Texas at Dallas

Dxg163330@utdallas.edu

# Abstract

Driven by ever-growing enterprise data, companies are facing huge expenses adding capacity to their existing Data Warehouses (DW). To decrease this spend, companies are looking at lower-cost alternatives that preserve their existing reporting and analytics. Data Warehouse Expansion using big data reduces warehouse management costs by enabling organizations to collect and retain data that was previously too expensive to store---data can now be made available for analysis and improved business insights at 1/10th to 1/50th the cost on a per terabyte basis. In addition, identifying and offloading infrequently used data from the existing data warehouse to low-cost Big Data stores yields immediate performance and cost benefits. This paper primarily based upon consideration of different solution and parameter and then comparing different data warehouse expansion method to prove the importance of big data.

1. **Introduction**

To study the comparison very rigorously and meticulously, the case study of one medium sized company have been taken. The company uses Teradata for the data warehouse environment for all structured data needs. However, the Teradata environment has already surpassed 10 TB and the company does not intend to grow it beyond 15 TB. There are few business units asking for data playground (sandbox) areas for exploration and discovery needs. In addition, the company would like analyze the web logs from company’s website. The web logs are currently stored in tape and are approximately 40 TB in size. The company management is looking for ways to analyze the web logs and provide the users with a ‘data discovery environment, which should not in the Production Teradata environment.

To propose best solution, the comparison study is carried out between data warehouse system expansion without Hadoop and data warehouse system expansion with Hadoop(80%) using complex total cost of ownership(TCO) and return on investment(ROI) for 6 years using World Wide Technology Data Warehouse Cost Saving Calculator. Various parameter have been taken into consideration and discussed in details for the comparison of two.

The new model architecture of the expanded data warehouse systems has been created to find out total hardware requirements and total cost of ownership for the 3 years. A comparative study has been carried out to compare cloud based big data implementation and premises based big data implementation using Microsoft Azure TCO calculator.

Various Hadoop contributors have been compared to choose best player using comparision of different parameter like File system, License Cost, Long term goal, consistency in Market, Partnership of Cloudera and Hortonworks with Teradata, etc.

After that conclusion/Executive summary have been derived to conclude which is the best system to implement and reasons for it.

1. **Hybrid Approach: Traditional data warehouse plus Hadoop**
   1. **Comparing different Parameters**

There are many ways in which data warehouse system without Hadoop and data warehouse system with Hadoop (80%) can be compared. The main questions is raised here that can Hadoop replace the traditional data warehouse systems. This question can be answered by considering following parameters.

**Processing Structured Data** is something that traditional database is already very good at. After all, structured data, by definition is easy to enter, store, query and analyze. It conforms nicely to a fixed schema model of neat columns and rows that can be manipulated by Structured Query Language (SQL) to establish relationships. As such, using Hadoop to process structured data would be comparable to running simple errands with a Formula One racecar. However, with the rise of big data, many of those simple errands have become quite complex, calling for a more powerful and streamlined solution than the data warehouse can offer.

**Storing, managing and analyzing massive volumes of semi-structured and unstructured data** is what Hadoop was purpose-built to do. Unlike structured data, found within the tidy confines of records, spreadsheets and files, semi-structured and unstructured data is raw, complex, and pours in from multiple sources such as emails, text documents, videos, photos, social media posts, Twitter feeds, sensors and clickstreams.

Most data warehouses are built on specialized infrastructure, processing large batches of data can be very costly. The other problem with RDBMS with respect to performing more complex big data workloads is that the raw data must first be put through a cleaning and structuring process called ETL (Extract, Transform, and Load) before putting it into the warehouse. But this pre-processing of data can be plagued with errors. Plus it permanently eliminates all potentially valuable raw data while limiting how the resultant “clean and simple” data can be queried.

[Hadoop as a Service](https://www.qubole.com/resources/webinars/deploying-hadoop-bare-metal-cloud/) provides a scalable solution to meet ever-increasing data storage and processing demands that the data warehouse can no longer handle. With its unlimited scale and on-demand access to compute and storage capacity, Hadoop as a Service is the perfect match for [big data processing](https://www.qubole.com/resources/webinars/spot-utilization/). Using tools found within the Hadoop ecosystem, such as Pig, Spark, Presto and others, Hadoop as a Service will help company to obtain the deeper insights often hidden in unstructured data that can propel business forward.

**Running constant and predictable workloads** is what existing data warehouse has been all about. And as a solution for meeting the demands of structured data—data that can be entered, stored, queried, and analyzed in a simple and straightforward manner—the data warehouse will continue to be a viable solution. But when it comes to handling massive volumes of unstructured data, that’s where the warehouse falls short.

**Running fluctuating workloads** to meet growing big data demands requires a scalable infrastructure that allows servers to be provisioned as needed. That’s where cloud-based Hadoop service comes in handy. With the ability to spin virtual servers up or down on demand within minutes, Hadoop in the cloud provides the flexible scalability company will need to handle fluctuating workloads.

**Keeping costs down** is a concern for every business in today’s ultra-competitive arena. And traditional relational databases are certainly cost effective. While on-premise Hadoop implementations save money by combining open source software with commodity servers, a cloud-based Hadoop platform will save you even more by eliminating the expense of physical servers and warehouse space entirely. Hybrid systems, which integrate cloud-based Hadoop with traditional relational databases, are fast gaining popularity as cost-effective ways for companies to leverage the benefits of both platforms.

**Running large distributed workloads** that address every file in the database is something that Hadoop handles very well, but not very fast. And so the tradeoff with this type of processing is slower time-to-insight.

**Shorter time-to-insight** necessitates interactive querying via the analysis of smaller data sets in near or real-time. And that’s a task that the data warehouse has been well equipped to handle. However, thanks to a powerful Hadoop processing engine called Spark, Hadoop—and in particular Hadoop as a Service—can handle both batch and streaming workloads at lightening fast speeds. Spark is designed for advanced, real-time analytics and has the framework and tools to deliver when shorter time-to-insight is critical.

The combination of a traditional database with cloud-based Hadoop platform can be a powerful, cost-effective analytical tool for business. Properly implemented, this hybridized data infrastructure allows companies to reap the benefits of both platforms by running small, highly interactive workloads in the data warehouse while using Hadoop to process very large and complex data sets to obtain deeper insights and drive competitive advantage.

* 1. **Comparative Study**

As part of our hybrid solution, comparative study has been conducted for 5 years and total cost per year and total saving per year have been calculated.

Following is some assumption and parameter which have been taken into consideration for the cost calculations.

**Size of DW(TB):** Current of size of data ware house System in TB. In our case, current size of Data warehouse system is 10 TB and we have more of 40 TB of weblogs data.So, total size of desired DW system will be 50 TB. Hence, making this parameter 50 TB.

**DW cost per TB($)**: It is assumed to be $20000

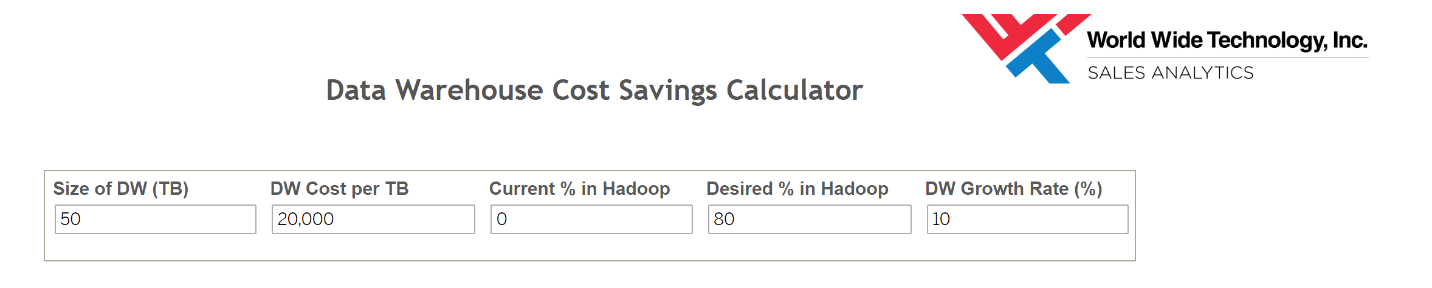
**Current % in Hadoop:** In our case, it is assumed that we don’t have any big data system implemented in data warehouse system and hence making this parameter 0%.

**Desired % in Hadoop:** From above study, after knowing the advantages of hybrid system, we are doing to implement the hybrid data ware house system having 80% of Hadoop share.Hence making this parameter as 80%

**DW Growth Rate:** it is assumed that the data ware house growth rate is 10% year by year for our company.

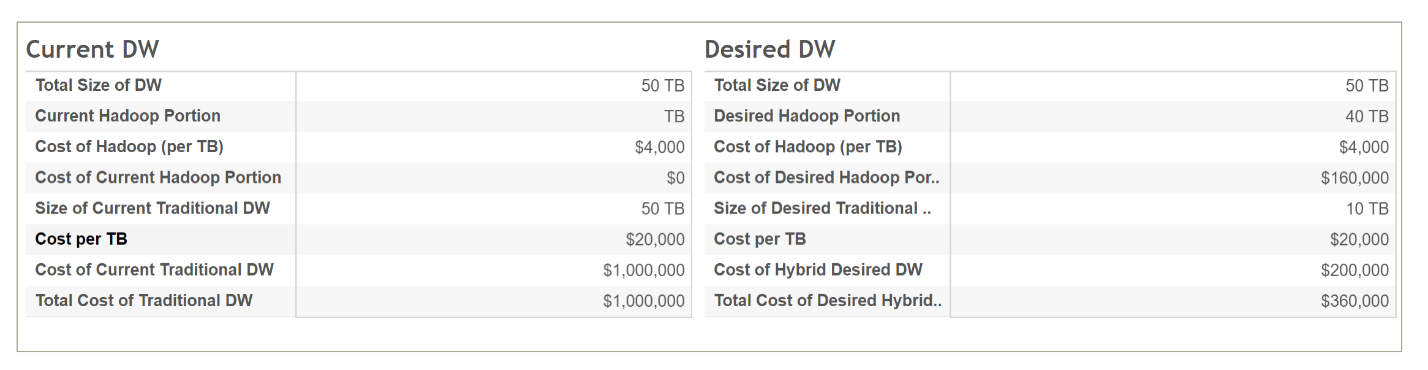
**Calculations :** <https://www2.wwt.com/dwo-calculator/>

**Input data:**

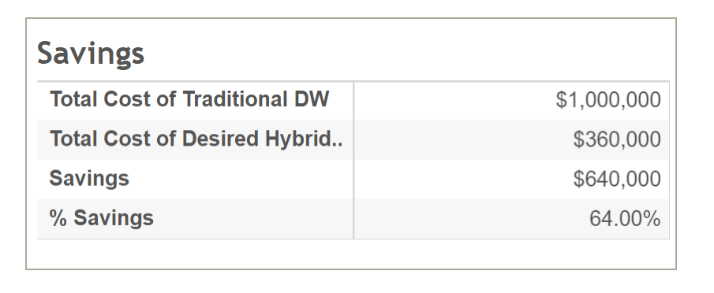


Following are the results:

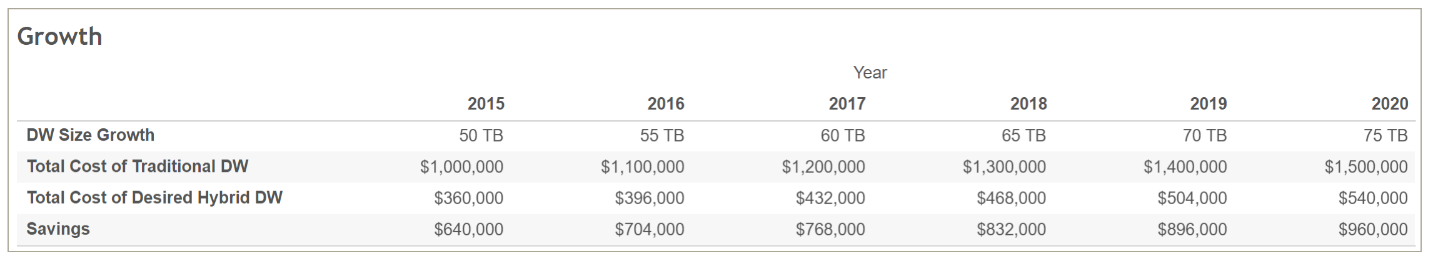
**Total cost of data ware house system without Hadoop vs With 80% Hadoop**



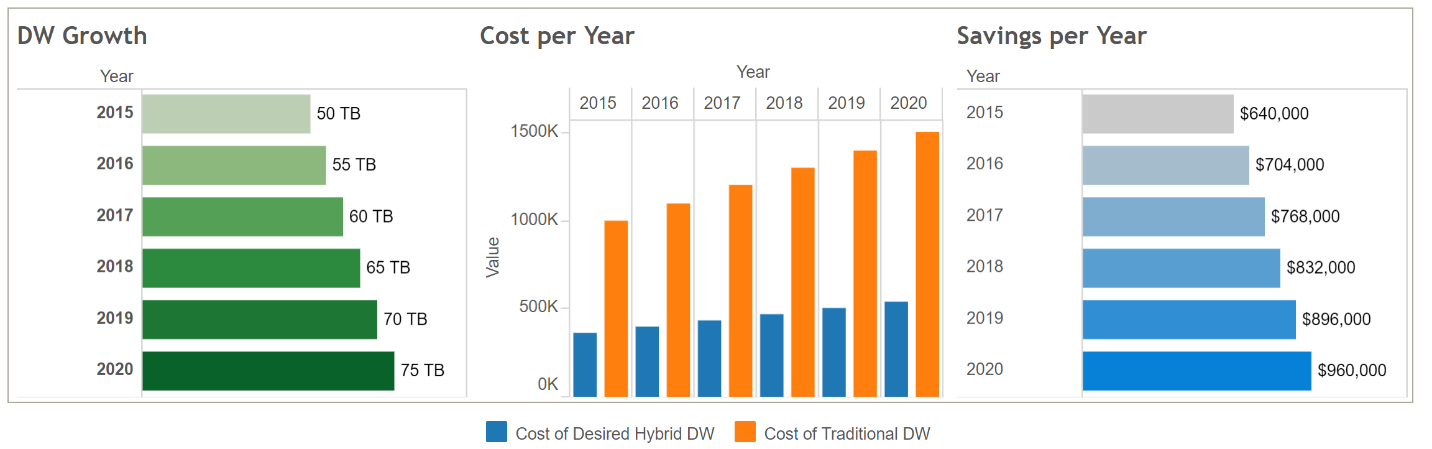
**Total saving: 64%**



**Growth Chart of cost Over the years:**

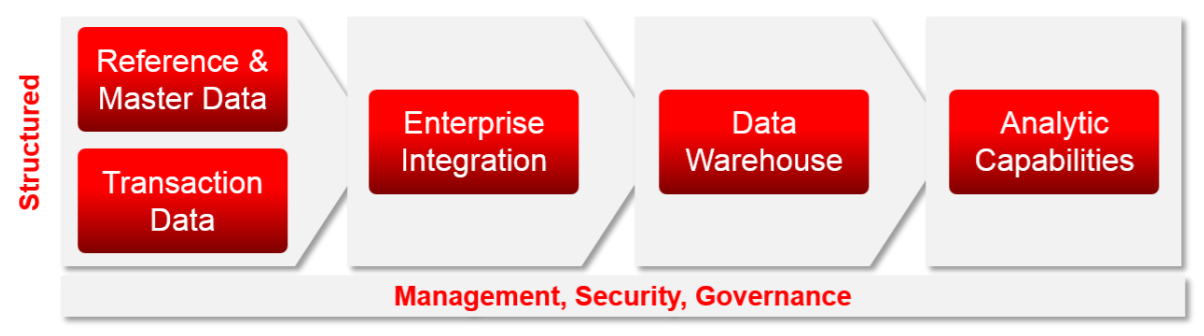


**Year wise cost and saving:**



It is quite apparent from above study that implementing hybrid data ware house system with 80% Hadoop system with have total saving of $960000 over six year timeframe. I will definitely recommend the company to go with hybrid data warehouse system with 80% of Hadoop portion.

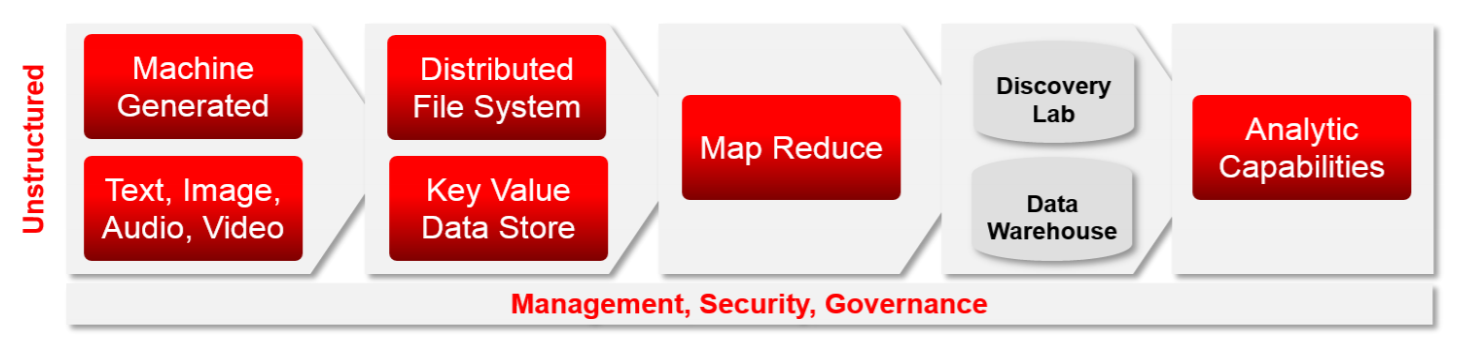
1. **Future Technical Architecture for the Hybrid data warehouse system with Hadoop**
   1. **Traditional Information Architecture Capabilities:** To understand the high-level architecture aspects of future Big Data plus existing warehouse, let’s understand present data warehouse system architecture for structured data of our model company. In the illustration, two data sources that use integration (ELT/ETL/Change Data Capture) techniques to transfer data into a DBMS data warehouse or operational data store, and then offer a wide variety of analytical capabilities to reveal the data. Some of these analytic capabilities include: dashboards, reporting, EPM/BI applications, summary and statistical query, semantic interpretations for textual data, and visualization tools for high-density data. In addition, some organizations have applied oversight and standardization across projects, and perhaps have matured the information architecture capability through managing it at the enterprise level.



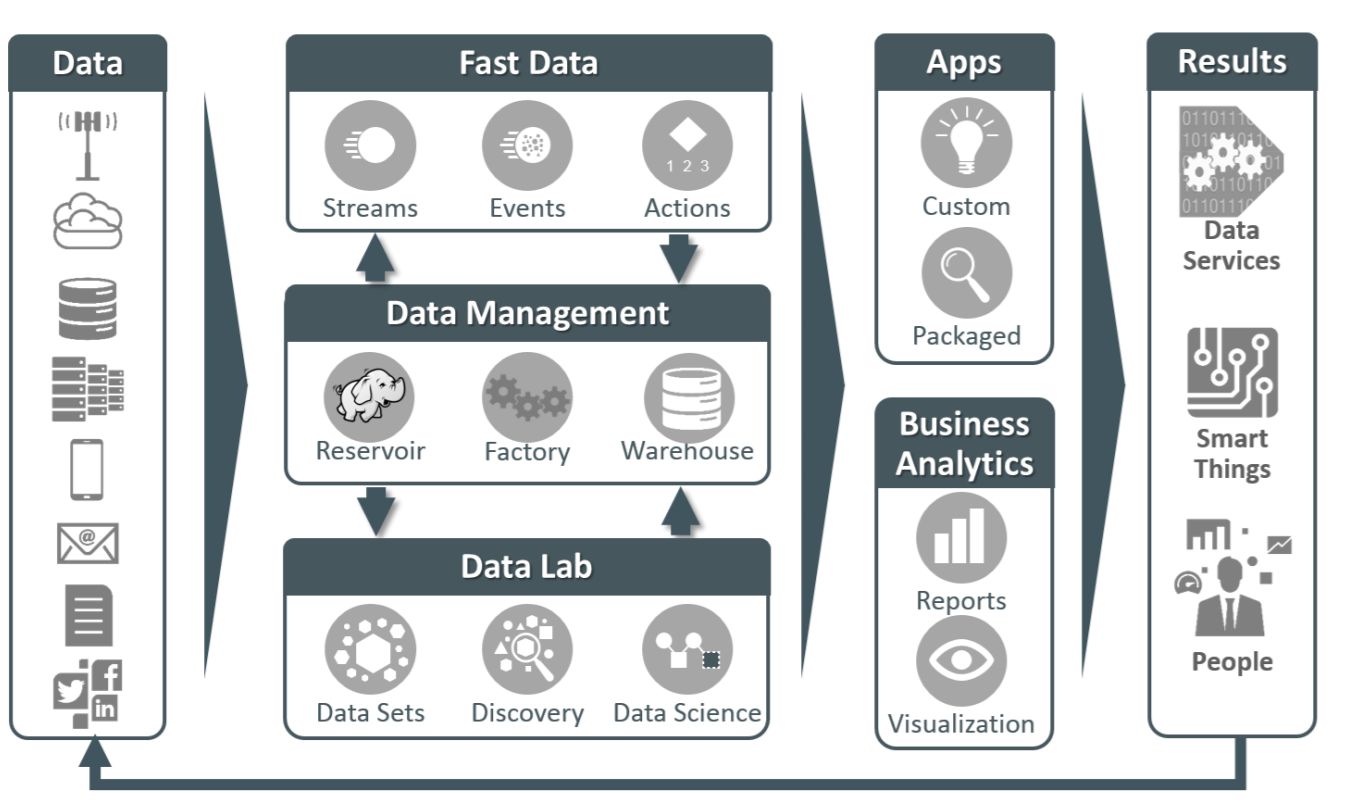
The key information architecture principles include treating data as an asset through a value, cost, and risk lens, and ensuring timeliness, quality, and accuracy of data. And, the enterprise architecture oversight responsibility is to establish and maintain a balanced governance approach including using a center of excellence for standards management and training.

* 1. **Adding Big Data Capabilities**

Below image illustrate future diagram of our model dataware house system architecture adding big data(Hadoop).The defining processing capabilities for big data architecture are to meet the volume, velocity, variety, and value requirements. Unique distributed (multi-node) parallel processing architectures have been created to parse these large data sets. There are differing technology strategies for real-time and batch processing storage requirements. For real-time, key-value data stores, such as NoSQL, allow for high performance, index-based retrieval. For batch processing, a technique known as “Map Reduce,” filters data according to a specific data discovery strategy. After the filtered data is discovered, it can be analyzed directly, loaded into other unstructured or semi-structured databases, sent to mobile devices, or merged into traditional data warehousing environment and correlated to structured data



* 1. **A Full Detailed Unified Reference Architecture**

Below image illustrates key components and flows and highlights the emergence of the Data Lab and various forms of new and traditional data collection. 

A description of these primary components:

**Fast Data:** Components which process data in-flight (streams) to identify actionable events and then determine next-best-action based on decision context and event profile data and persist in a durable storage system. The decision context relies on data in the data reservoir or other enterprise information stores.

**Reservoir:** Economical, scale-out storage and parallel processing for data which does not have stringent requirements for data formalization or modelling. Typically manifested as a Hadoop cluster or staging area in a relational database. » Factory: Management and orchestration of data into and between the Data Reservoir and Enterprise Information Store as well as the rapid provisioning of data into the Discovery Lab for agile discovery. Warehouse: Large scale formalized and modelled business critical data store, typically manifested by a Data Warehouse or Data Marts.

**Data Lab:** A set of data stores, processing engines, and analysis tools separate from the data management activities to facilitate the discovery of new knowledge. Key requirements include rapid data provisioning and sub setting, data security/governance, and rapid statistical processing for large data sets.

**Business Analytics:** A range of end user and analytic tools for business Intelligence, faceted navigation, and data mining analytic tools including dashboards, reports, and mobile access for timely and accurate reporting.

**Apps:** A collection of prebuilt adapters and application programming interfaces that enable all data sources and processing to be directly integrated into custom or packaged business applications.

1. **Construction of Model Project**

we will need a model project, and here are the assumptions for our theoretical medium-sized, enterprise-class big data project:

* 1. **Assumptions**

**Users**: 150 end-users: 100 business end-users, 25 data scientists/analysts.

Analytics Consumption: One-third ad-hoc, one-third daily, one-third monthly.

**Servers**: Enterprise-grade with newer chipsets, backup power supply, high storage density (to ensure, for example, that Hadoop doesn't become storage-limited), plenty of cores to support parallelization with more cores in the name node, plus plenty of memory to handle complex queries and columnar-based analytics.

**Storage**: Total raw data capacity of 384 terabytes (TB), in a balanced mix of refreshes (i.e., monthly, weekly, daily, real-time). Note that in the example in Table 1, the math suggests more terabytes (4 nodes \* 96 TB/node = 384 TB), but one has to account for replication (4x replication in Hadoop for example), peaks, compression, and sharding to arrive at the actual usable capacity. We will use a usability rate of 25%, which in this case, rounded up, yields about 96 TB.

**Network**: Dedicated and fast bandwidth, with multiple InfiniBand switches to deal with contention; there can be varying amounts of replication and data movement within a cluster, but a big data infrastructure needs to ensure it doesn't become network-bound.

**Queries**: A diverse set of activities that spans from simple select statements to complex joins.

Integration and Information Management: Five new points of integration/transformation; in the process of design, additional data sources will be added in addition to, for example, a data warehouse. Those sources will require integration/transformation and information management work, and related licenses and hardware.

* 1. **Hardware/Software Cost Calculations**

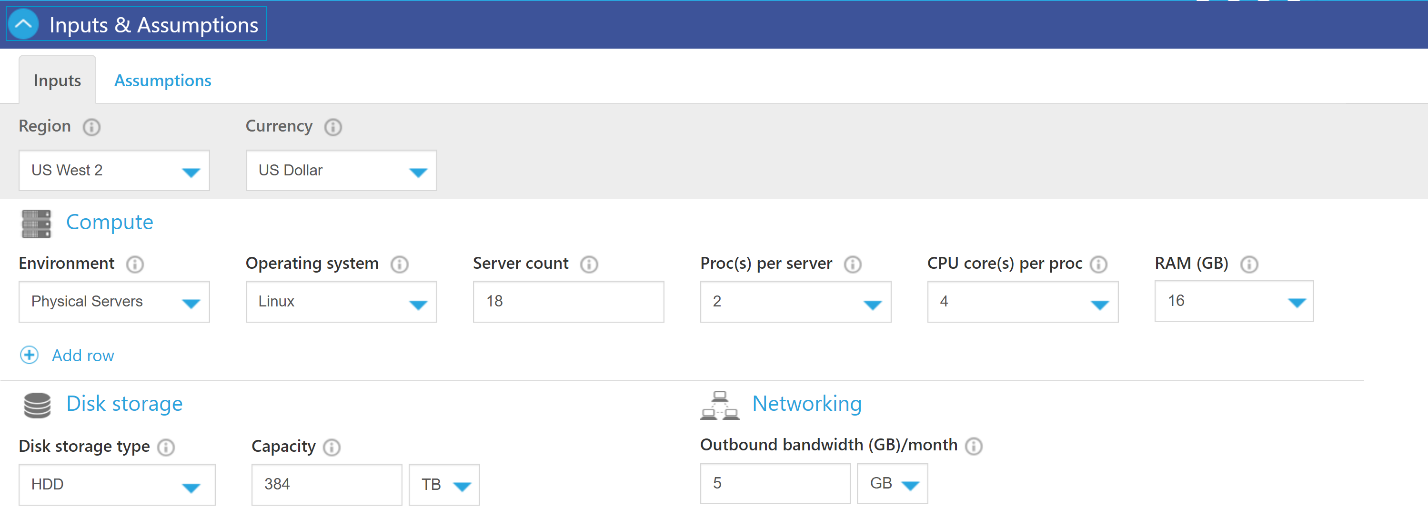
|  |  |  |
| --- | --- | --- |
| Item | Cost ($) | Notes |
| Servers | 157,460 | Four @ $39,365 each; enterprise-class, each with: two Intel Xeon 18-core processors, eight 16GB memory kits, twelve 8TB hard drives, network card, and dual power supplies |
| Networking | 27,000 | Two Mellanox 36-port Infiniband switches at $10,500 each, plus $6,000 cabling |
| Rack enclosure | 2,500 | Including cables, looms, patch panels, etc. |
| Hardware Support | 28044 | @15% of list cost of above items |
| OS Licence | 4892 | Red Hat Enterprise Linux at $1,223 annually for each of 4 nodes for 3 years |
| Hadoop Licensing | 28,800 | Cloudera Enterprise Data Hub at $7,200 annually for each of 4 nodes for 3 years |
| Total Hardware/Software Cost | 248,696 |  |

* 1. **Total Cost of Ownership Calculations**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Acquisition Costs** | **Operating Costs** | **Total($)** | **% of** |
| **TCO** |
| **Software** | 33692 | 0 | **33692** | 6.47% |
| **Hardware** | 157460 | 28044 | **185504** | 35.63% |
| **Personnel** | 0 | 270,000 | **270000** | 51.85% |
| **NW & Comm** | 27000 | 0 | **27000** | 5.19% |
| **Facilities** | 2500 | 2000 | **4500** | 0.86% |
| **Total ($)** | **220,652** | **300,044** | **520,696** | — |
| **% of TCO** | 42.38% | 57.62% | — | 100.00% |

1. **Hadoop: On premises Vs Cloud Comparison**

The comparative study has been conducted using Microsoft Azure TCO calculator(https://www.tco.microsoft.com/Home/Calculator) for the cloud based services to compare two different approaches for our model system:



* 1. **Input Assumption/Parameter**

**Environment**: we want to replace our physical servers with cloud base services. For that this parameter will be set as physical servers.

**Operating system :** Our model as linux as operating system on each server and hence putiing this parameter as linux.

Our system as total 4 servers ( 2 processor / server and 18 core/processor) and hence we have total (144 cores) and hence we are dividing our cores in Azure system as follows. **(because azure is not allowing to out the server configuration designed for our system )**

**So ,making another system with same number of cores (144)**

**Server count:** 18

**Proc(s) per server:** 2

**Cpu core(s) per proc:** 4

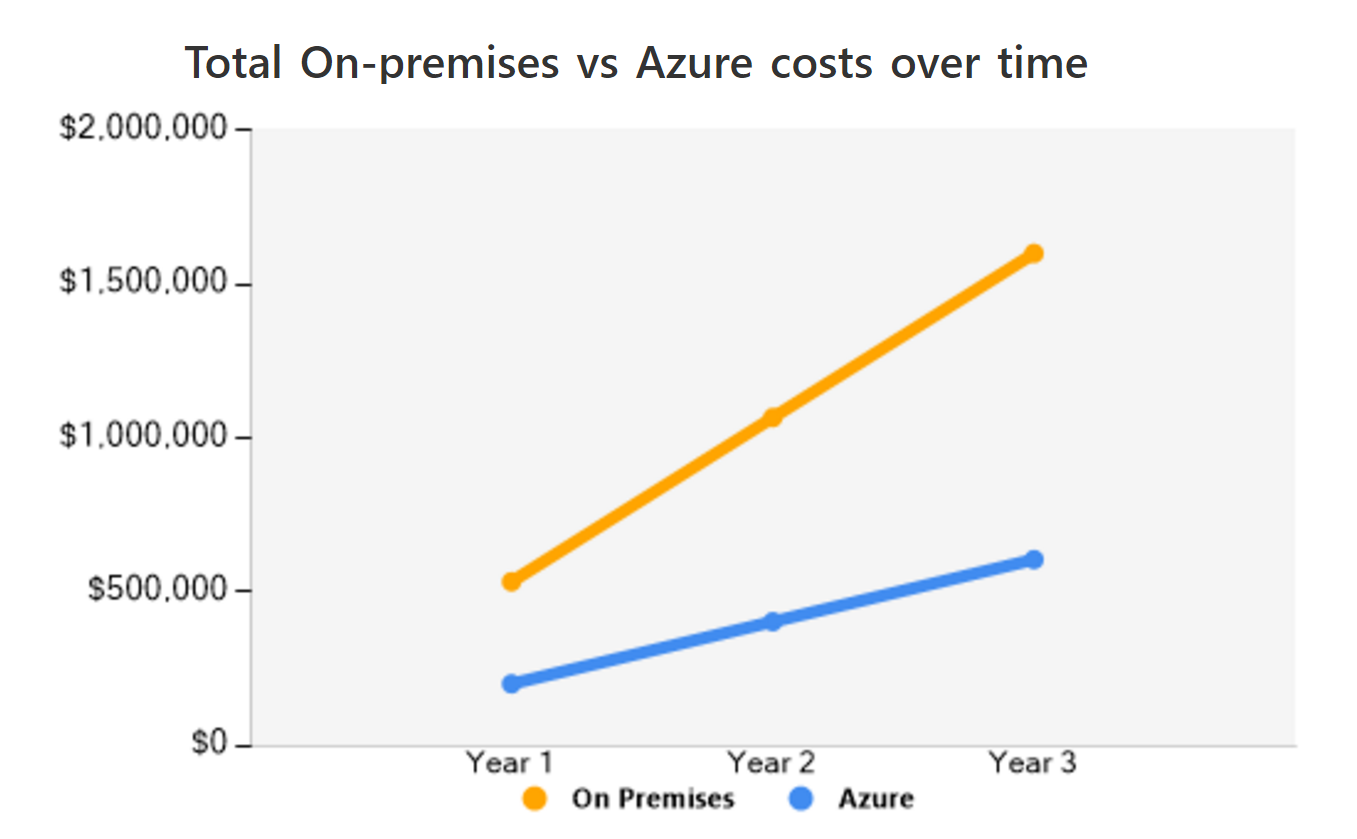
**RAM per server:** 18 GB

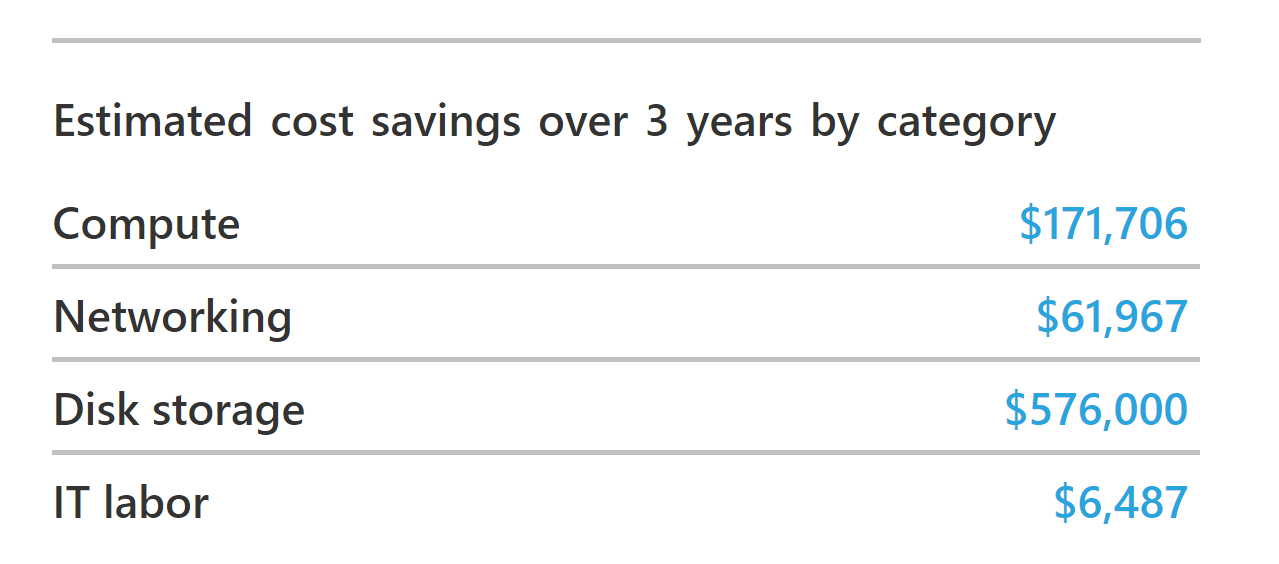
**Disk storage type:** HDD

**Capacity:** we want to have total 384 TB system and hence putting this values as 384 TB

Networking: outbound bandwidth – 5 GB

* 1. **Results**





|  |  |  |  |
| --- | --- | --- | --- |
| On-Premises costs breakdown summary | |  | |
| Category | | Costs | |
| Compute | | $258,761 | |
| Hardware | | $206,467 | |
| Software | | $0 | |
| Electricity | | $15,620 | |
| Data center | | $36,674 | |
| Networking | | $61,976 | |
| Disk storage | | $1,267,200 | |
| IT labor | | $13,953 | |
| Total costs | | $1,601,890 | |
|  | |  | |
| Azure costs breakdown summary |  | |  | |
| Category | Web Direct costs | | CPP costs with EA Pricing | |
| Compute | $135,329 | | $87,055 | |
| Networking | $0 | | $9 | |
| Disk storage | $691,200 | | $511,488 | |
| IT labor | $7,466 | | $7,466 | |
| Total | $833,995 | | $606,018 | |

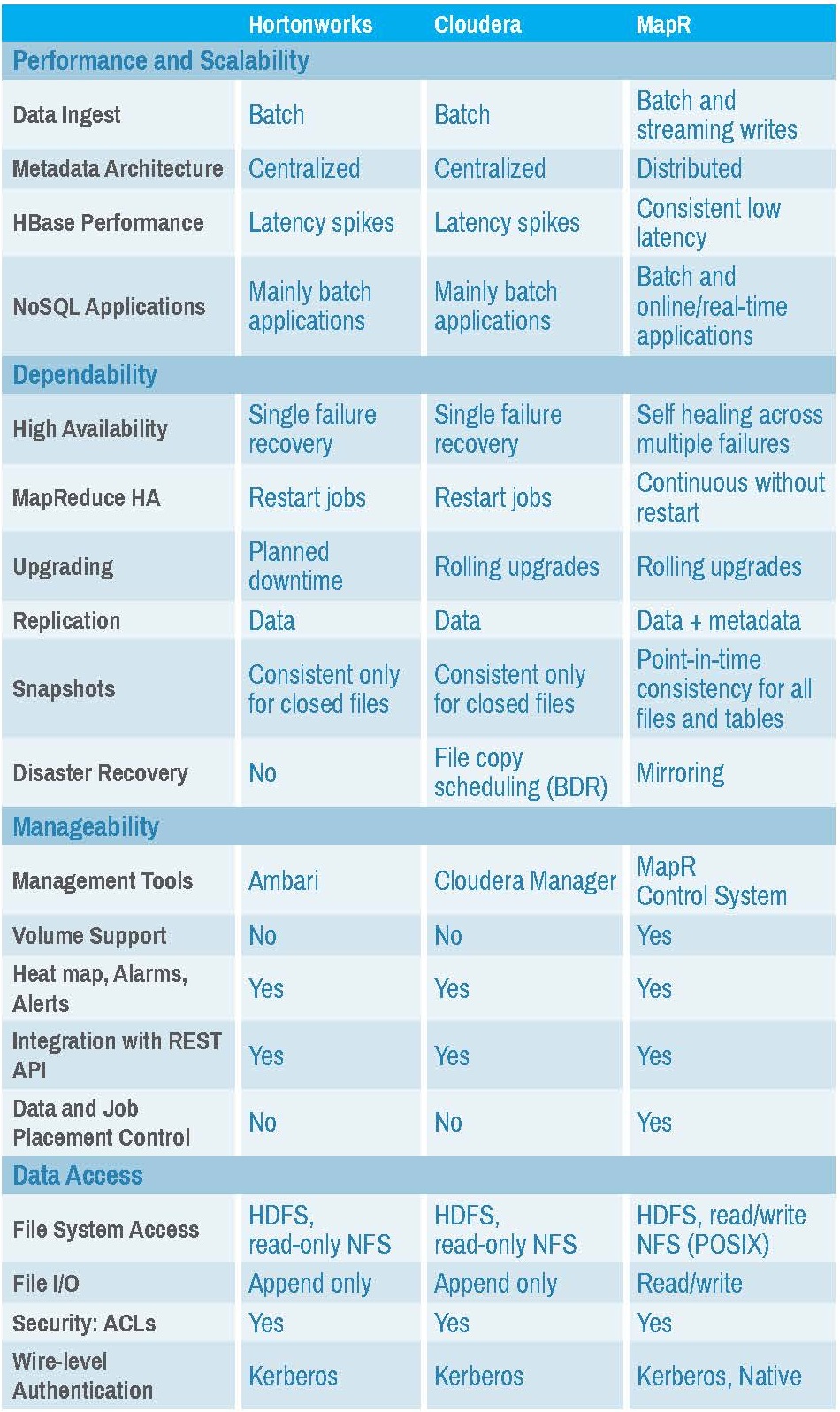
Estimated savings are **$ 995,872 (62%) over 3 years with Microsoft Azure(cloud based)**

**Conclusion:** It is certain from the above model study that cloud based services are far cheaper then the actual premise based software services and hence I will recommend to go with the cloud based services.

1. **Selection of best Hadoop distributor**

It is very importance to select best Hadoop distributer over the time as it will affect the overall ROI of the total data warehouse system.

Cloudera, Hortonworks and MapR are three most popular Hadoop distributers in the market.following the executive summary of all feature of



Comparison Parameter which is taken into considerations:

**File system:** Both Cloudera and Hortonworks uses HDFS has its basic file system. MapR replaces HDFS component and instead uses its own proprietary file system, called MapRFS. MapRFS helps incorporate enterprise-grade features into Hadoop, enabling more efficient management of data, reliability and most importantly, ease of use. However, it is very uncommon to have different file system then usual file system and hence we are eliminating the MapR from our consideration.

**Cloudera Teradata Partnership:**

Cloudera and Teradata—the respective leaders in Hadoop and data warehousing—have joined forces in the areas of technology integrations and solutions, support, and trusted advisor consultation.

* Teradata Appliance for Hadoop with Cloudera is a purpose-built, integrated hardware and software solution for data at scale. The appliance runs Cloudera Enterprise with the additional benefits of simplified manageability and enterprise grade reliability that comes from an engineered appliance, all backed by Teradata and Cloudera’s world-class customer support. By minimizing the number of moving parts required for deployment and operations, the appliance allows companies to achieve faster time-to-value—thereby leveraging your current investments in technology and resources.
* Teradata’s Unified Data Architecture—a big data architecture that includes Cloudera Enterprise, Teradata, and Teradata Aster Discovery Platform—enables clients to orchestrate workloads in such a way as to realize the right set of capabilities and economics for the right job. This architecture comes with a set of robust software technologies—Teradata QueryGrid™, Teradata Loom®, Teradata Connector for Hadoop, and Teradata Aster File Store—to provide a self-service platform with seamless access to all your data.
* Teradata Customer Support provides maintenance and service  for Cloudera Enterprise customers who  require that a single vendor  support the full ecosystem.

Teradata Professional Services for Hadoop and data warehouse are easily deployed for Cloudera Enterprise—with skills such as data cleansing, governance, data lineage, predictive analytics, and security. Teradata’s recent acquisition of Think Big brings strong Hadoop and Cloudera-specific skills to mutual customers.

**Hortonworks Teradata Partnership:**

Hortonworks and Teradata believe that no single analytic system can meet all customers’ needs, and leading data driven organizations will deploy an analytical ecosystem. Harnessing relational and non-relational data, and running SQL and non-SQL analytics requires specialized technologies. That is why Hortonworks and Teradata have a longstanding engineering partnership focused on delivering the analytical ecosystem to the enterprise through best in class analytic engines and co-engineering of software to orchestrate the processing and strip out the complexity typically associated with a hybrid environment.

**Proprietary management software:** Cloudera has a proprietary management software Cloudera Manager, SQL query handling interface Impala, as well as Cloudera Search for easy and real-time access of products. Hortonworks has no proprietary software, uses Ambari for management and Stinger for handling queries, and Apache Solr for searches of data. This is very strong point that we have to consider.

**License**: Cloudera has a commercial license, while Hortonworks has open source license. Cloudera also allows the use of its open- source projects free of cost, but the package doesn’t include the management suite Cloudera Manager or any other proprietary software.

**Long term goal:** Cloudera has announced that its long term goal is to become an enterprise data hub, thus diminishing the need of data warehouse. Hortonworks, on the other hand, remains firmly a provider of Hadoop distro, and has partnered with data warehousing company Teradata. By this point Cloudera is more favorable then Horton works for our model project of dataware house expansion.

**Consistency in Market:** Cloudera has been the oldest player in the market, with more than 350 customers. But Hortonworks is fast catching up and has made more innovations in the Hadoop ecosystem in the recent past. Cloudera has several enterprise softwares overlaid on its open source distributions to aid the consumers, whereas Hortonworks strives to provide a framework comprising only of open source projects.

From all of the above points it can be concluded that for our model data warehouse project cloudera turns out to be clear winner and hence I will strongly recommend to go with cloudera distributor for the Hadoop installation.

1. **Executive Summary/Conclusion**

In initial case, the company uses Teradata for the data warehouse environment for all structured data needs. However, the Teradata environment has already surpassed 10 TB and the company does not intend to grow it beyond 15 TB. There are few business units asking for data playground (sandbox) areasThe web logs are currently stored in tape and are approximately 40 TB in size. The company management is looking for ways to analyze the web logs and provide the users with a ‘data discovery environment.

To propose best solution, the comparison study is carried out between data warehouse system expansion without Hadoop and data warehouse system expansion with Hadoop(80%) using complex total cost of ownership(TCO) and return on investment(ROI) for 6 years. that implementing hybrid data ware house system with 80% Hadoop system with have total saving of $960000 over six year timeframe which is total 64% Saving. So, It is strongly recommended that company must include Hadoop(big data) into data warehouse expansion to get more saving over the time.

The new model architecture of the expanded data warehouse systems has been created. In this new model, we have also included big data system along with new features like data labs which will provide the facility of data sandbox which was asked by some business units for the discovery and expansion. Data discovery systems were already present in the old architecture, its compatibility with new big data system has been added in our new architecture.

The model of having additional 384 TB memory ( four – 2 processor 18- core servers) in hadoop system has been designed and the total hardware and software cost, Maintenance, support, license cost have been calculated for the new Hadoop solution and its total cost of ownership(TCO) has been calculated over 3 years, which turns out to be $520,696.

A comparative study has been carried out to compare cloud based big data implementation and premises based big data implementation using Microsoft Azure TCO calculator. As a result, it has been concluded that company can have savings of 62% of premise base cost over 3 years with cloud based Hadoop installation. So, it is highly recommended that company should go for cloud base Hadoop installation rather then on premises Hadoop installations.

Various Hadoop contributors have been compared to choose best player using comparision of different parameter like File system, License Cost, Long term goal, consistency in Market, Partnership of Cloudera and Hortonworks with Teradata, etc. Hence by considering all parameter, company should go with cloudera as Hadoop distributor because of its outstanding capabilities.

1. **References**
   1. http://www.cio.com/article/3005705/software/calculating-the-total-cost-of-ownership-for-enterprise-software.html
   2. <http://www.axia-consulting.co.uk/html/basic_roi_calculation.html>
   3. <https://www.cognizant.com/whitepapers/from-relational-database-management-to-big-data-solutions-for-data-migration-testing-codex1439.pdf>
   4. <http://www.information-management.com/news/big-data-analytics/Data-Lakes-Data-Warehouse-Big-Data-10027291-1.html>
   5. <https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf>
   6. <https://vision.cloudera.com/start-reducing-tco-and-accelerating-time-to-value-of-big-data/>
   7. <http://www.datacenterknowledge.com/archives/2013/04/01/building-a-cloud-savvy-model-for-tco-and-roi/>
   8. <http://www.cloudera.com/content/dam/www/static/documents/whitepapers/Accenture%20Cloudera%20SAP%20Insight_Powered_Enterprise_POV_FINAL.pdf>
   9. <http://info.hortonworks.com/rs/549-QAL-086/images/hortonworks-data-architecture-optimization.pdf>
   10. <https://www.hpe.com/h20195/v2/GetDocument.aspx?docname=4AA6-8931ENW&doctype=Technical%20white%20paper&doclang=EN_US&searchquery=&cc=us&lc=en>
   11. <https://www.hpe.com/h20195/v2/GetDocument.aspx?docname=4AA6-3305ENW&doctype=Technical%20white%20paper&doclang=EN_US&searchquery=&cc=us&lc=en>
   12. <https://blogs.oracle.com/datawarehousing/entry/updated_price_comparison_for_big>
   13. <http://www.oracle.com/us/technologies/big-data/eng-systems-for-big-data-esg-wp-2852701.pdf>
   14. <https://www-01.ibm.com/software/info/middleware/db2/uk/pdf/mmr-pco.pdf>
   15. <https://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/>
   16. <https://www.dezyre.com/article/cloudera-vs-hortonworks-vs-mapr-hadoop-distribution-comparison-/190>
   17. <https://www.quora.com/What-distribution-should-I-choose-Cloudera-Hortonworks-or-MapR>
   18. <http://www.americandigital.com/cloudera-hortonworks-and-mapr-comparing-the-top-three-hadoop-distributions/>
   19. <https://www2.wwt.com/dwo-calculator/>
   20. <https://www.qubole.com/blog/big-data/hadoop-and-data-warehouse/>
   21. <http://data-magnum.com/cloudera-vs-hortonworks-vs-mapr-has-mapr-already-won-this-contest/>
   22. <https://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/>
   23. <https://www.cloudera.com/partners/solutions/teradata.html>
   24. <http://hortonworks.com/partner/teradata/>
   25. <https://www.emc.com/collateral/emc-perspective/h10915-ep-pdf-data-warehouse-modernization.pdf>
   26. <http://www.tcs.com/SiteCollectionDocuments/White-Papers/BFS-Whitepaper-Big-Data-Warehousing-0313-1.pdf>
   27. <https://www.ibm.com/developerworks/library/ba-augment-data-warehouse1/index.html>
   28. https://www.tco.microsoft.com/Home/Calculator