Big Data Capacity Planning: Achieving Right Sized Hadoop Clusters and Optimized Operations

Abstract

Businesses are considering more opportunities to leverage data for different purposes, impacting resources and resulting in poor loading and response times. Hadoop is increasingly being adopted across industry verticals for information management and analytics. In addition to new business related capabilities, it offers a host of options for IT simplification and cost reduction. Initiatives such as offloads are at the heart of this type of optimization. As a result, Hadoop capacity planning should be carried out as the first step in both IT-driven and business-driven use cases whenever Big Data projects are considered.

Understanding Big Data and Its Defining Characteristics

Hadoop or Big Data ecosystem offer a set of techniques and technologies that include new forms of integration capabilities to detect the hidden value of large, diverse, and complex datasets. The data is generated from various enterprise sources, sensors, social media posts, digital pictures and videos, purchase transaction records, and cell phone GPS signals. In simpler terms, Big Data is defined by characteristics such as volume, velocity, variety, and veracity.

Leveraging Hadoop to Solve the Big Data Challenges

Hadoop enables storage and processing of large amounts of data without investing in expensive, proprietary hardware. It facilitates distributed and limitless parallel processing of huge amounts of data across inexpensive, industry-standard commodity servers that store as well as process the data. Hadoop's unlimited scalability allows organizations to store data without worrying about performance, storage costs, archival, and retention periods. Its key advantages include scalability, high fault tolerance and low upfront costs. It also facilitates quick analysis of massive collections of records without requiring the data to be first modeled, cleansed, and loaded.

Big Data capacity planning takes a wide variety of aspects into consideration. This includes incoming data volumes, data to be retained, types of data, methods by which the data arrives, and forecast of needs volumes. It also includes aspects such as data aggregates used for building analytics based on this data, type of hardware needed, frequency of processing, incoming data intervals, and whether the cluster is intended for batch processing or inmemory capability is required for tools such as Impala.

Different methods by which Hadoop capacity planning for Big Data projects include:

Enabling Efficient Capacity Planning for Hadoop Clusters

The Hadoop cluster capacity planning methodology addresses workload characterization and forecasting. Here, workload characterization refers to how MapReduce jobs interact with the storage layers and forecasting addresses prediction of future data volumes for processing and storage.

Commonly, Hadoop clusters are sized based on data storage, data volumes processed by a job, data types, and response time required. Large quantities of data require more systems to process the same. Adding new nodes to the cluster brings in more computing resources in addition to new storage capacity. The sizing of a cluster comes from the specifics of a workload which include CPU workload, memory, storage, disk I/O and network bandwidth.

For high efficiency, the Hadoop Distributed Files System (HDFS) should have high throughput hard drives with an underlying file system that supports the HDFS read and write pattern. HDFS works well with one big read or write at a time, with block sizes of 64MB, 128MB, 256MB, 512MB, and all the way up to 1GB. This should also be supported by a network layer that is fast enough to cope with intermediate data transfer.

Key Considerations	Recommendations
How is the data ingested, and at what frequency?	The architecture needs to be planned based on the ingestion type (in streams, batches or from an RDBMS system) and supported with capacity planning.
Does the Hadoop system need to be read- or write-intensive?	If the Hadoop system to be developed is write intensive, resources necessary to quickly complete the writes need to be planned. A few distributions can be leveraged to write one copy and confirm that it is done, while the others write all three copies (replication factor three) and confirm that the replication is done. If it is to be read intensive, necessary memory (perhaps in-memory) and network resources should be increased.
How many concurrent users will there be access the system?	If the number of users is large, it is advisable to increase the nodes and their resources (RAM).
Latency - How quickly is the data to be accessed? (Will batch processing suffice or is faster processing expected?)	If data is to be processed and accessed quickly, in-memory architecture needs to be planned.

Data and system related aspects to be considered during capacity planning

Projecting Required Big Data Capacity

We start with 1 TB of daily data from Year 1 and assume 15% data growth per quarter. Further, assuming a 15% year-on-year growth in data volumes and 1,080 TB of data in Year 1, by the end of Year 5 the capacity may grow to 8,295 TB of data. If we were to assume a 30% year-on-year growth in data volumes and 1080 TB of data in Year 1, then by the end of Year 5, the capacity might grow to 50,598 TB of data.

The following formula can be used to estimate Hadoop storage and arrive at the required number of data nodes:

Hadoop Storage (H) = C*R*S/(1-i)

Legend

C: Average compression ratio

R: Replication factor

S: Size of data to be moved to Hadoop

i: Intermediate factor

Estimating Required Hadoop Storage and Number of Data Nodes

With no compression, C equals 1. The replication factor is assumed to be 3 and the intermediate factor 0.25 or ¼. The calculation for H in this case becomes:

H= 1*3*S/(1-(1/4)) = 3*S/(3/4) = 4*S

The required Hadoop storage in this instance is estimated to be four times the initial data size.

The following formula can be used to estimate the number of data nodes:

(n) = H/D = C*R*S/(1-i)/D

D: Disk space available per node

Let us assume that 8 TB is the available disk space per node, each node comprising 10 disks of 1 TB capacity each, minus 2 disks for operating system. Also, assuming the initial data size to be 600 TB:

N = 600/8 = 75

Thus, 75 data nodes are needed in this case.

If complex processing is anticipated, then it is recommended to have at least 10% additional vacant space to accommodate such processing. This 10% is an addition to the 20% set aside for OS installation and operation.

The memory needed for each node can be calculated as follows:
Total memory needed = [(memory per CPU core) * (number of CPU's core)] + data node process memory + data node task tracker memory + OS memory

Each data node will comprise a number of data blocks on the cluster. As a thumb rule, it should be ensured that an increase in the number of data nodes is supported by a corresponding increase in the RAM as well.

Facilitating Effective Hardware Configuration for Hadoop Clusters

Unlike traditional systems that fetch data from databases and process it in application servers, the Hadoop framework sends the processing logic to each data node in the cluster that stores and processes the data in parallel. The cluster of these balanced machines should thus satisfy data storage and processing requirements. It is also imperative to take the replication factor into consideration during capacity planning to ensure fault tolerance and data reliability. Network resources play a vital role while executing jobs and reading and writing to the disks over the network.

The following elements need to be taken into consideration while building a Hadoop cluster:

- Namenode (and secondary namenode)
- Job tracker (resource manager)
- Task tracker (node manager)
- Data node

Additional Recommendations to Improve Capacity Planning

Additional recommendations that can be implemented to ensure efficient capacity planning are:

- While computing memory requirements, it is advisable to dedicate 10% of the memory for the Java Virtual Machine, required to process programs such as MapReduce.
- Hadoop should be configured with strict heap size restrictions to avoid memory swapping to the disk. Swapping impacts the performance of the MapReduce job. This can also be avoided by configuring data node machines with more RAM and setting appropriate kernel settings on Linux distribution.
- While planning the capacity, additional components such as HBase, Impala and Search may be taken into consideration as they run on the data node process to maintain data locality

Conclusion

With new digital technologies gaining greater prominence, it is imminent that Big Data with its quantitative abilities will lay the foundation for improved qualitative analysis. The expected increase in data implies an ever increasing focus on capacity planning, a critical requirement for all production systems. Capacity planning is an exercise and a continuous practice to arrive at the right infrastructure that caters to the current, near future, and future needs of a business.

Businesses that embrace capacity planning will realize the ability to efficiently handle massive amounts of data and manage the user base. This in turn has the potential to positively impact the bottom line and help organizations gain a competitive edge in the marketplace.

References

This can include citations and references

About The Author

Rajasekhar Reddy Pentareddy

Rajasekhar Reddy Pentareddy has over 12 years of IT experience spanning multiple technologies and domains. His areas of expertise include Big Data, Data Warehousing and Business Intelligence.

Contact

Visit TCS' Analytics & Insights unit page for more information

Email: analytics.insights@tcs.com

Blog: Digital Reimagination

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled, infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

TCS Design Services I M I 12 I 16