UNIT I

INTRODUCTION TO BIG DATA

EVOLUTION OF BIG DATA

It all starts with the explosion in the amount of data we have generated since the dawn of the digital age. This is largely due to the rise of computers, the Internet and technology capable of capturing data from the world we live in. Going back even before computers and databases, we had paper transaction records, customer records etc. Computers, and particularly spreadsheets and databases, gave us a way to store and organize data on a large scale. Suddenly, information was available at the click of a mouse.

We‘ve come a long way since early spreadsheets and databases, though. Today, every two days we create as much data as we did from the beginning of time until 2000. And the amount of data we‘re creating continues to increase rapidly.

Nowadays, almost every action we take leaves a digital trail. We generate data whenever we go online, when we carry our GPS-equipped smartphones, when we communicate with our friends through social media or chat applications, and when we shop. You could say we leave digital footprints with everything we do that involves a digital action, which is almost everything. On top of this, the amount of machine-generated data is rapidly growing too.

How does Big Data work?

Big Data works on the principle that the more you know about anything or any situation, the more reliably you can gain new insights and make predictions about what will happen in the future. By comparing more data points, relationships begin to emerge that were previously hidden, and these relationships enable us to learn and make smarter decisions. Most commonly, this is done through a process that involves building models, based on the data we can collect, and then running simulations, tweaking the value of data points each time and monitoring how it impacts our results. This process is automated – today‘s advanced analytics technology will run millions of these simulations, tweaking all the possible variables until it finds a pattern – or an insight – that helps solve the problem it is working on.

Anything that wasn‘t easily organised into rows and columns was simply too difficult to work with and was ignored. Now though, advances in storage and analytics mean that we can capture, store and work with different types of data. Thus, ―data‖ can now mean anything from databases to photos, videos, sound recordings, written text and sensor data.

To make sense of all this messy data, Big Data projects often use cutting-edge analytics involving artificial intelligence and machine learning. By teaching computers to identify what this data represents– through image recognition or natural language processing, for example – they can learn to spot patterns much more quickly and reliably than humans.

APEC

Industrial impact of Big Data in 2020:

Machine Learning and Artificial Intelligence will proliferate

The deadly duo will get beefed up with more muscles. Continuing with our round-up of the latest trends in big data, we will now take stock of how AI and ML are doing in the big data industry. Artificial intelligence and machine learning are the two sturdy technological workhorses working hard to transform the seemingly unwieldy big data into an approachable stack. Deploying them will enable businesses to experience the algorithmic magic via various practical applications like video analytics, pattern recognition, customer churn modelling, dynamic pricing, fraud detection, and many more. IDC predicts that spending on AI and ML will rise to $57.6 billion in 2021. Similarly, companies pouring money into AI are optimistic that their revenues will increase by 39% in 2020.

Raise of Quantum Computing

The next computing juggernaut is getting ready to strike, the quantum computers. These are the powerful computers that have principles of Quantum Mechanics working on their base. Although, you must wait patiently for at least another half a decade before the technology hits the mainstream. One thing is for sure; it will push the envelope of traditional computing and do analytics of unthinkable proportions. Predictions for big data are thus incomplete without quantum computing

Edge analytics will gain increased traction

The phenomenal proliferation of IoT devices demands a different kind of analytics solution and edge analytics is probably the befitting answer. Edge analytics means conducting real-time analysis of data at the edge of a network or the point where data is being captured without transporting that data to a centralized data store. For its on-site nature, it offers certain cool benefits: reduction in bandwidth requirements, minimization of the impact of load spikes, reduction in latency, and superb scalability. Surely, edge analytics will find more corporate takers in future. One survey says between 2017 and 2025, the total edge analytics market will expand at a moderately high CAGR of 27.6% to pass the $25 billion mark. This will have a noticeable impact on big data analytics as well.

Dark data

So, what is Dark Data, anyway? Every day, businesses collect a lot of digital data that is stored but is never used for any purposes other than regulatory compliance and since we never know when it might become useful. Since data storage is easier, businesses are not leaving anything out. Old data formats, files, documents within the organization are just lying there and being accumulated in huge amounts every second. This unstructured data can be a goldmine of insights, but only if it is analysed effectively. According to IBM, by 2020, upwards of 93% of all data will fall under Dark Data category. Thus, big data in 2020 will inarguably reflect the inclusion of Dark Data. The fact is we must process all types of data to extract maximum benefit from data crunching.

Usage:

This ever-growing stream of sensor information, photographs, text, voice and video data means we can now use data in ways that were not possible before. This is revolutionising the world of business across almost every industry. Companies can now accurately predict what specific segments of customers will want to buy, and when to buy. And Big Data is also helping companies run their operations in a much more efficient way.

Even outside of business, Big Data projects are already helping to change our world in several ways, such as:

- **Improving healthcare:** Data-driven medicine involves analysing vast numbers of medical records and images for patterns that can help spot disease early and develop new medicines.
- **Predicting and responding to natural and man-made disasters:** Sensor data can be analysed to predict where earthquakes are likely to strike next, and patterns of human behavior give clues that help organisations give relief to survivors and much more.
- **Preventing crime**: Police forces are increasingly adopting data-driven strategies based on their own intelligence and public data sets in order to deploy resources more efficiently and act as a deterrent where one is needed.
- **Marketing effectiveness:** Big Data, along with being able to help businesses and organizations in making smart decisions also drastically increases the sales and marketing effectiveness of the businesses and organizations thus highly improving their performances in the industry.
- **Prediction and Decision making:** Now that the organizations can analyse Big Data, they have successfully started using Big Data to mitigate risks, revolving around various factors of their businesses. Using Big Data to reduce the risks regarding the decisions of the organizations and making predictions has become one of the many benefits coming from big data in industries.

Concerns:

Big Data gives us unprecedented insights and opportunities, but it also raises concerns and questions that must be addressed:

- **Data privacy:** The Big Data we now generate contains a lot of information about our personal lives, much of which we have a right to keep private
- **Data security:** Even if we decide we are happy for someone to have our data for a purpose, can we trust them to keep it safe?
- **Data discrimination:** When everything is known, will it become acceptable to discriminate against people based on data we have on their lives? We already use credit scoring to decide who can borrow money, and insurance is heavily data-driven.
- **Data quality:** Not enough emphasis on quality and contextual relevance. The trend with technology is collecting more raw data closer to the end user. The danger is data in raw format has quality issues. Reducing the gap between the end user and raw data increases issues in data quality.

APEC

Facing up to these challenges is an important part of Big Data, and they must be addressed by organisations who want to take advantage of data. Failure to do so can leave businesses vulnerable, not just in terms of their reputation, but also legally and financially.

BEST PRACTICES FOR BIG DATA ANALYTICS

Business is awash in data—and also big data analytics programs meant to make sense of this data and apply it toward competitive advantage. A recent Gartner study found that more than 75 percent of businesses either use big data or plan to spin it up within the next two years.

Not all big data analytics operations are created equal, however; there‘s plenty of noise around big data, but some big data analytics initiatives still don‘t capture the bulk of useful business intelligence and others struggling getting off the ground.

For those businesses currently struggling with the data, or still planning their approach, here are five best practices for effectively using big data analytics.

1. Start at the End

The most successful big data analytics operations start with the pressing questions that need answering and work backwards. While technology considerations can steal the focus, utility comes from starting with the problem and figuring out how big data can help find a solution.

There are many directions that most businesses can take their data, so the best operations let key questions drive the process and not the technology tools themselves.

–Businesses should not try to boil the ocean, and should work backwards from the expected outcomes,‖ says Jean-Luc Chatelain, chief technology officer for Accenture Analytics, part of Accenture Digital.

2. Build an Analytics Culture

Change management and training are important components of a good big data analytics program. For greatest impact, employees must think in terms of data and analytics so they turn to it when developing strategy and solving business problems. This requires a considerable adjustment in both how employees and businesses operate.

Training also is key so employees know how to use the tools that make sense of the data; the best big data system is useless if employees can‘t functionally use it.

–We approach big data analytics programs with the same mindset as any other analytic or transformational program: You must address the people, process and technology in the organization rather than just data and technology,‖ says Paul Roma, chief analytics officer for Deloitte Consulting.

–Be ready to change the way you work,‖ adds Luc Burgelman, CEO of NGDATA, a firm that helps financial services, media firms and telecoms with big data utilization. –Big data has the power to transform your entire business but only if you are flexible and prepared to be open to change.

3. Re-Engineer Data Systems for Analytics

An increasing range and volume of devices now generate data, creating substantial variation both in sources and types of data. An important component of a successful big data analytics program is re-engineering the data pipelines so data gets to where it needs to be and in a form that is useful for analysis. Many existing systems were not developed for today‘s big data analysis needs.

–This is still an issue in many businesses, where the data supply chain is blocked or significantly more complex than is necessary, leading to _trapped data‘ that value can‘t be extracted from,‖ says Chatelain at Accenture Digital. –From a data engineering perspective, we often talk about re-architecting the data supply chain, in part to break down silos in where data is coming from, but also to make sure insights from data are available where they are relevant.‖

4. Focus on Useful Data Islands

There‘s a lot of data. Not all of it can be mined and fully exploited. One key of the most successful big data analytics operations is correctly identifying which islands of data offer the most promise.

–Finding and using precise data is rapidly becoming the Holy Grail of analytics activities,‖ says Chatelain. ―Enterprises are taking action to address the challenges present in grappling with big data, but [they] continue to struggle to identify the islands of relevant data in the big data ocean.‖

Burgelman at NGDATA also stresses the importance of data selection.

–Most companies are overwhelmed by the sheer volume of the data they possess, much of which is irrelevant to the stated goal at hand and is just taking up space in the database,‖ he says.
–By determining which parameters will have the most impact for your company, you‘ll be able to make better use of the data you have through a more focused approach rather than attempting to sort through it all.‖

5. Iterate Often

Business velocity is at an all-time high thanks to more globally connected markets and rapidly evolving information technology. The data opportunities are constantly changing, and with that comes the need for an agile, iterative approach toward data mining and analysis. Good big data analytics systems are nimble and always iterating as new technology and data opportunities emerge.

Big data itself can help drive this evolution.

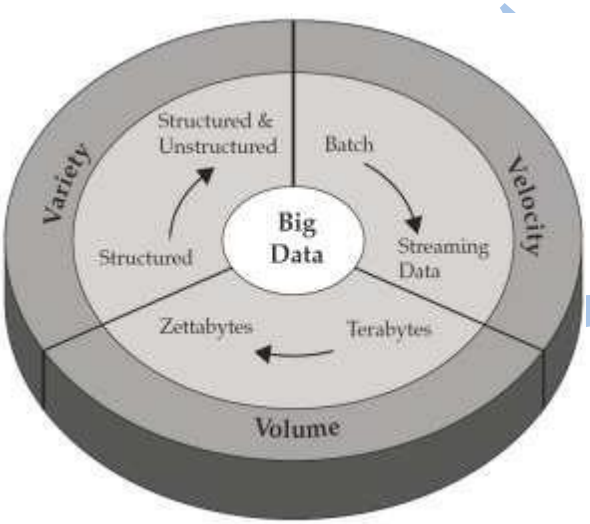
–One of the amazing things about big data analytics is that it can help organizations gain a better understanding of what they don‘t know,‖ says Burgelman. –So as data comes in and conclusions are reached, you‘ve got to be flexible and open to changing the scope of the project. Don‘t be afraid to ask new questions of your data on an ongoing basis.‖

The importance of effective big data use grows by the day. This makes analytics best practices all the more important, and these five top the list.

BIG DATA CHARACTERISTICS

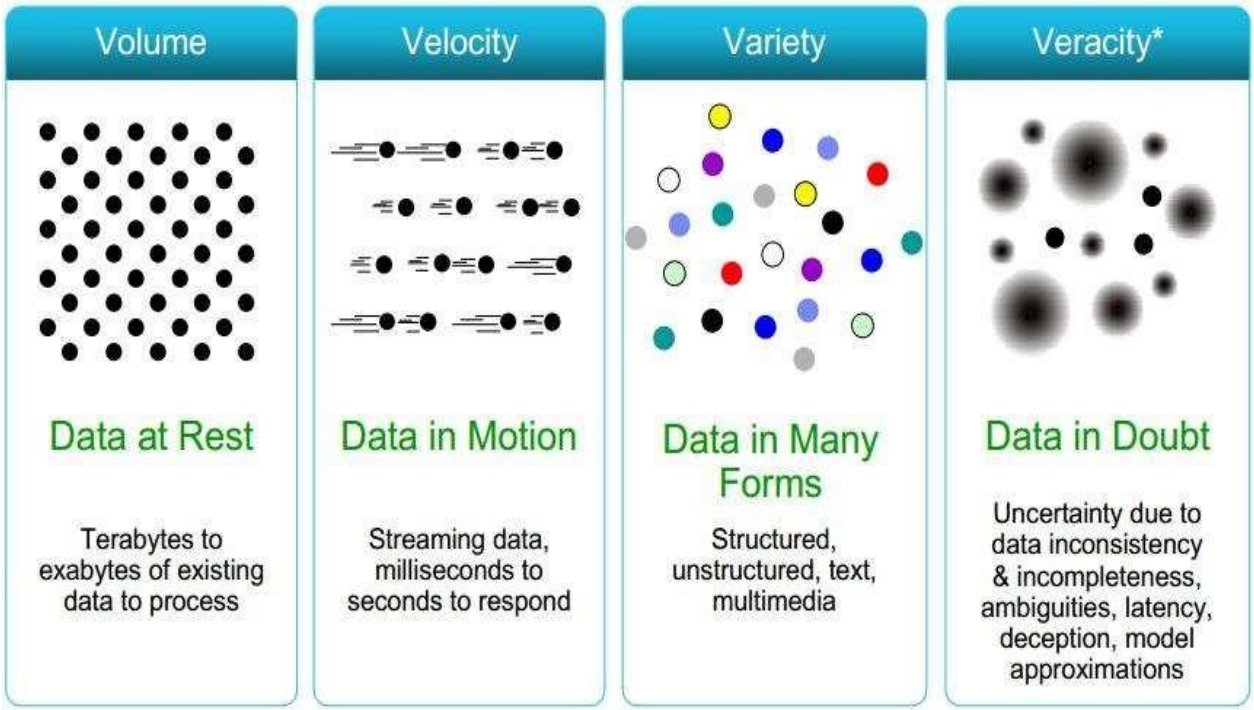
Three attributes stand out as defining Big Data characteristics:

1. **Volume**
   Huge volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
2. **Variety**
   Complexity of data types and structures: Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
3. **Velocity**
   Speed of new data creation and growth: Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.



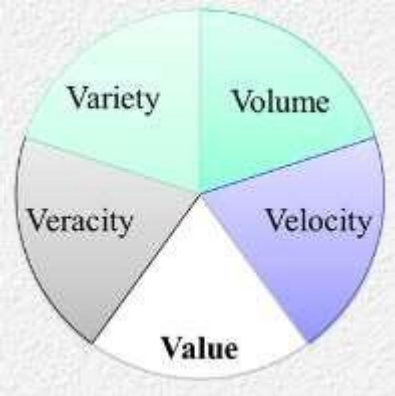
This can be extended to add a fourth V called Veracity of data(Data in Doubt)

4. **Veracity**



5. Value

There is another V to take into account when looking at big data: Value. Having access to big data is no good unless we can turn it into value. Companies are starting to generate amazing value from their big data.



VALIDATING (AGAINST) THE HYPE: ORGANIZATIONAL FITNESS

Even as the excitement around big data analytics reaches a fevered pitch, it remains a technology-driven activity. And as we speculated in Chapter 1, there are a number of factors that need to be considered before making a decision regarding adopting that technology. But all of those factors need to be taken into consideration; just because big data is feasible within the organization, it does not necessarily mean that it is reasonable.

Unless there are clear processes for determining the value proposition, there is a risk that it will remain a fad until it hits the disappointment phase of the hype cycle. At that point, hopes may be dashed when it becomes clear that the basis for the investments in the technology was not grounded in expectations for clear business improvements.

As a way to properly ground any initiatives around big data, one initial task would be to evaluate the organization‘s fitness as a combination of the five factors presented in Chapter 1: feasibility, reasonability, value, integrability, and sustainability. Table 1.1 provides a sample framework for determining a score for each of these factors ranging from 0 (lowest level) to 4 (highest level).

The resulting scores can be reviewed (an example of a radar chart is shown in Figure 1.1). Each of these variables is, for the most part, somewhat subjective, but there are ways of introducing a degree of objectivity, especially when considering the value of big data.

Table 2.1 Quantifying Organizational Readiness

Score by Dimension	0	1	2	3	4
Feasibility	Evaluation of new technology is not officially sanctioned	Organization tests new technologies in reaction to market pressure	Organization evaluates and tests new technologies after market evidence of successful use	Organization is open to evaluation of new technology Adoption of technology on an *ad hoc* basis based on convincing business justifications	Organization encourages evaluation and testing of new technology Clear decision process for adoption or rejection Organization supports allocation of time to innovation
Reasonability	Organization's resource requirements for near-, mid-, and long-terms are satisfactorily met	Organization's resource requirements for near- and mid-terms are satisfactorily met, unclear as to whether long-term needs are met	Organization's resource requirements for near-term is satisfactorily met, unclear as to whether mid- and long-term needs are met	Business challenges are expected to have resource requirements in the mid- and long-terms that will exceed the capability of the existing and planned environment	Business challenges have resource requirements that clearly exceed the capability of the existing and planned environment Organization's go-forward business model is highly information-centric
Value	Investment in hardware resources, software tools, skills training, and ongoing management and maintenance exceeds the expected quantifiable value	The expected quantifiable value widely is evenly balanced by an investment in hardware resources, software tools, skills training, and ongoing management and maintenance	Selected instances of perceived value may suggest a positive return on investment	Expectations for some quantifiable value for investing in limited aspects of the technology	The expected quantifiable value widely exceeds the investment in hardware resources, software tools, skills training, and ongoing management and maintenance
Integrability	Significant impediments to incorporating any nontraditional technology into environment	Willingness to invest effort in determining ways to integrate technology, with some successes	New technologies can be integrated into the environment within limitations and with some level of effort	Clear processes exist for migrating or integrating new technologies, but require dedicated resources and level of effort	No constraints or impediments to fully integrate technology into operational environment

Table 2.1 (Continued)

Score by Dimension	0	1	2	3	4
Sustainability	No plan in place for acquiring funding for ongoing management and maintenance costs No plan for managing skills inventory	Continued funding for maintenance and engagement is given on an *ad hoc* basis Sustainability is at risk on a continuous basis	Need for year-by-year business justifications for continued funding	Business justifications ensure continued funding and investments in skills	Program management office effective in absorbing and amortizing management and maintenance costs Program for continuous skills enhancement and training

THE PROMOTION OF THE VALUE OF BIG DATA

That being said, a thoughtful approach must differentiate between hype and reality, and one way to do this is to review the difference between what is being said about big data and what is being done with big data. A scan of existing content on the ‑value of big data‖ sheds interesting light on what is being promoted as the expected result of big data analytics and, more interestingly, how familiar those expectations sound. A good example is provided within an economic study on the value of big data (titled ‑Data Equity—Unlocking the Value of Big Data‖), undertaken and published by the Center for Economics and Business Research (CEBR) that speaks to the cumulative value of:

• optimized consumer spending as a result of improved targeted customer marketing;
• improvements to research and analytics within the manufacturing sectors to lead to new product development;
• improvements in strategizing and business planning leading to innovation and new start-up companies;
• predictive analytics for improving supply chain management to optimize stock management, replenishment, and forecasting;
• improving the scope and accuracy of fraud detection.

Curiously, these are exactly the same types of benefits promoted by business intelligence and data warehouse tools vendors and system integrators for the past 15_20 years, namely:

• Better targeted customer marketing
• Improved product analytics
• Improved business planning
• Improved supply chain management
• Improved analysis for fraud, waste, and abuse

Further articles, papers, and vendor messaging on big data reinforce these presumptions, but if these were the same improvements promised by wave after wave of new technologies, what makes big data different?

BIG DATA USE CASES

The answer must lie in the ‑democratization‖ of high-performance capabilities, which is inherent in the characteristics of the big data analytics application development environment. This environment largely consists of a methodology for elastically harnessing parallel computing resources and distributed storage, scalable performance management, along with data exchange via high-speed networks.

The result is improved performance and scalability, and we can examine another data point that provides self-reported descriptions using big data techniques, namely, the enumeration of projects listed at The Apache Software Foundation‘s PoweredBy Hadoop Web site (http://wiki.apache.org/hadoop/PoweredBy).

A scan of the list allows us to group most of those applications into these categories:

• Business intelligence, querying, reporting, searching, including many implementation of searching, filtering, indexing, speeding up aggregation for reporting and for report generation, trend analysis, search optimization, and general information retrieval.
• Improved performance for common data management operations, with the majority focusing on log storage, data storage and archiving, followed by sorting, running joins, extraction/transformation/ loading (ETL) processing, other types of data conversions, as well as duplicate analysis and elimination.
• Non-database applications, such as image processing, text processing in preparation for publishing, genome sequencing, protein sequencing and structure prediction, web crawling, and monitoring workflow processes.
• Data mining and analytical applications, including social network analysis, facial recognition, profile matching, other types of text analytics, web mining, machine learning, information extraction, personalization and recommendation analysis, ad optimization, and behavior analysis.

In turn, the core capabilities that are implemented using the big data application can be further abstracted into more fundamental categories:
• Counting functions applied to large bodies of data that can be segmented and distributed among a pool of computing and storage resources, such as document indexing, concept filtering, and aggregation (counts and sums).
• Scanning functions that can be broken up into parallel threads, such as sorting, data transformations, semantic text analysis, pattern recognition, and searching.
• Modeling capabilities for analysis and prediction.
• Storing large datasets while providing relatively rapid access.
Generally, Processing applications can combine these core capabilities in different ways.

CHARACTERISTICS OF BIG DATA APPLICATIONS

What is interesting to note is that most of the applications reported by Hadoop users are not necessarily new applications. Rather, there are many familiar applications, except that the availability of a low-cost high-performance computing framework either allows more users to develop these applications, run larger deployments, or speed up the execution time. This, coupled with a further review of the different types of applications, suggests that of the limited scenarios discussed as big data success stories, the big data approach is mostly suited to addressing or solving business problems that are subject to one or more of the following criteria:

1. Data throttling: The business challenge has an existing solution, but on traditional hardware, the performance of a solution is throttled as a result of data accessibility, data latency, data availability, or limits on bandwidth in relation to the size of inputs.
1. Computation-restricted throttling: There are existing algorithms, but they are heuristic and have not been implemented because the expected computational performance has not been met with conventional systems.
3. Large data volumes: The analytical application combines a multitude of existing large datasets and data streams with high rates of data creation and delivery.
4. Significant data variety: The 11data in the different sources vary in structure and content, and some (or much) of the data is unstructured.
5. Benefits from data parallelization: Because of the reduced data dependencies, the application‘s runtime can be improved through task or thread-level parallelization applied to independent data segments.
So what, how does this relate to business problems whose solutions are suited to big data analytics applications? These criteria can be used to assess the degree to which business problems are suited to big data technology. As a prime example, ETL processing is hampered by data throttling and

computation throttling, can involve large data volumes, may consume a variety of different types of datasets, and can benefit from data parallelization. This is the equivalent of a big data ‒home run‖ application!

PERCEPTION AND QUANTIFICATION OF VALUE

So far we have looked at two facets of the appropriateness of big data, with the first being organizational fitness and the second being suitability of the business challenge. The third facet must also be folded into the equation, and that is big data‘s contribution to the organization. In essence, these facets drill down into the question of value and whether using big data significantly contributes to adding value to the organization by:

• Increasing revenues: As an example, an expectation of using a recommendation engine would be to increase same-customer sales by adding more items into the market basket.

• Lowering costs: As an example, using a big data platform built on commodity hardware for ETL would reduce or eliminate the need for more specialized servers used for data staging, thereby reducing the storage footprint and reducing operating costs.

• Increasing productivity: Increasing the speed for the pattern analysis and matching done for fraud analysis helps to identify more instances of suspicious behavior faster, allowing for actions to be taken more quickly and transform the organization from being focused on recovery of funds to proactive prevention of fraud

• Reducing risk: Using a big data platform or collecting many thousands of streams of automated sensor data can provide full visibility into the current state of a power grid, in which unusual events could be rapidly investigated to determine if a risk of an imminent outage can be reduced.

UNDERSTANDING BIG DATA STORAGE

As we have discussed in much of the book so far, most, if not all big data applications achieve their performance and scalability through deployment on a collection of storage and computing resources bound together within a runtime environment. In essence, the ability to design, develop, and implement a big data application is directly dependent on an awareness of the architecture of the underlying computing platform, both from a hardware and more importantly from a software perspective.

One commonality among the different appliances and frameworks is the adaptation of tools to leverage the combination of collections of four key computing resources:

1. Processing capability, often referred to as a CPU, processor, or node. Generally speaking, modern processing nodes often incorporate multiple cores that are individual CPUs that share the node‘s memory and are managed and scheduled together, allowing multiple tasks to be run simultaneously; this is known as multithreading.

1. Memory, which holds the data that the processing node is currently working on. Most single node machines have a limit to the amount of memory.

3. Storage, providing persistence of data—the place where datasets are loaded, and from which the data is loaded into memory to be processed.

4. Network, which provides the ‒pipes‖ through which datasets are exchanged between different processing and storage nodes. Because single-node computers are limited in their capacity, they cannot easily accommodate massive amounts of data. That is why the high-performance platforms are composed of collections of computers in which the massive amounts of data and requirements for processing can be distributed among a pool of resources.

A GENERAL OVERVIEW OF HIGH-PERFORMANCE ARCHITECTURE

Most high-performance platforms are created by connecting multiple nodes together via a variety of network topologies. Specialty appliances may differ in the specifics of the configurations, as do software appliances. However, the general architecture distinguishes the management of computing

resources (and corresponding allocation of tasks) and the management of the data across the network of storage nodes, as is seen in Figure 7.1.

In this configuration, a master job manager oversees the pool of processing nodes, assigns tasks, and monitors the activity. At the same time, a storage manager oversees the data storage pool and distributes datasets across the collection of storage resources. While there is no a priori requirement that there be any colocation of data and processing tasks, it is beneficial from a performance perspective to ensure that the threads process data that is local, or close to minimize the costs of data access latency.

To get a better understanding of the layering and interactions within a big data platform, we will examine the Apache Hadoop software stack, since the architecture is published and open for review. Hadoop is essentially a collection of open source projects that are combined to enable a software-based big data appliance. We begin with the core aspects of Hadoop's utilities, upon which the next layer in the stack is propped, namely Hadoop distributed file systems (HDFS) and MapReduce. A new generation framework for job scheduling and cluster management is being developed under the name YARN.

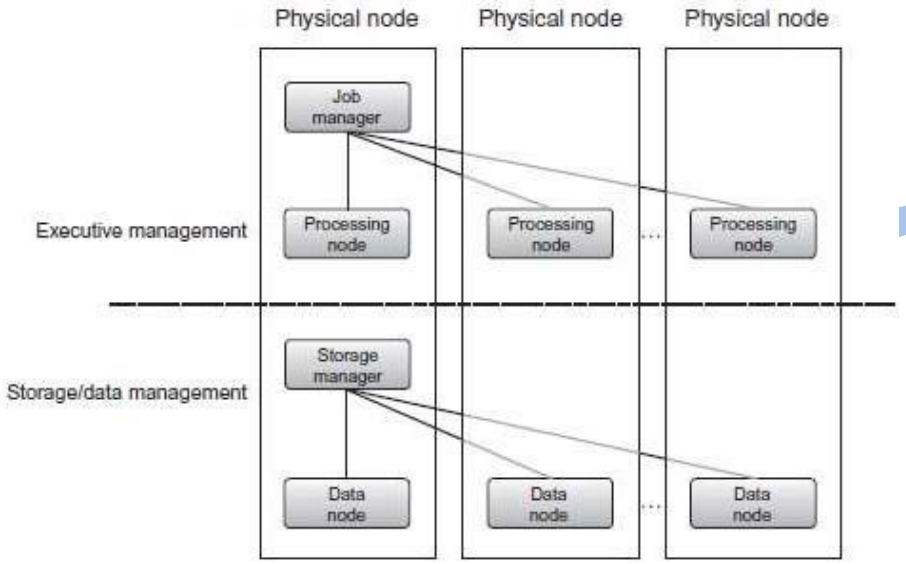


Figure 7.1 Typical organization of resources in a big data platform.

HDFS

HDFS attempts to enable the storage of large files, and does this by distributing the data among a pool of data nodes. A single name node (sometimes referred to as NameNode) runs in a cluster, associated with one or more data nodes, and provide the management of a typical hierarchical file organization and namespace. The name node effectively coordinates the interaction with the distributed data nodes.

The creation of a file in HDFS appears to be a single file, even though it blocks ‒chunks‖ of the file into pieces that are stored on individual data nodes.

The name node maintains metadata about each file as well as the history of changes to file metadata. That metadata includes an enumeration of the managed files, properties of the files, and the file system, as well as the mapping of blocks to files at the data nodes.

The data node itself does not manage any information about the logical HDFS file; rather, it treats each data block as a separate file and shares the critical information with the name node.

Once a file is created, as data is written to the file, it is actually cached in a temporary file. When the amount of the data in that temporary file is enough to fill a block in an HDFS file, the name node is alerted to transition that temporary file into a block that is committed to a permanent data node, which is also then incorporated into the file management scheme. HDFS provides a level of fault tolerance through data replication. An application can specify the degree of replication (i.e.,

the number of copies made) when a file is created. The name node also manages replication, attempting to optimize the marshaling and communication of replicated data in relation to the cluster‘s configuration and corresponding efficient use of network bandwidth. This is increasingly important in larger environments consisting of multiple racks of data servers, since communication among nodes on the same rack is generally faster than between server node sin different racks.
HDFS attempts to maintain awareness of data node locations across the hierarchical configuration.
In essence, HDFS provides performance through distribution of data and fault tolerance through replication. The result is a level of robustness for reliable massive file storage. Enabling this level of reliability should be facilitated through a number of key tasks for failure management, some of which are already deployed within HDFS while others are not currently implemented:
• Monitoring: There is a continuous ―heartbeat‖ communication between the data nodes to the name node. If a data node‘s heartbeat is not heard by the name node, the data node is considered to have failed and is no longer available. In this case, a replica is employed to replace the failed node, and a change is made to the replication scheme.
• Rebalancing: This is a process of automatically migrating blocks of data from one data node to another when there is free space, when there is an increased demand for the data and moving it may improve performance (such as moving from a traditional disk drive to a solid-state drive that is much faster or can accommodate increased numbers of simultaneous accesses), or an increased need to replication in reaction to more frequent node failures.
• Managing integrity: HDFS uses checksums, which are effectively ―digital signatures‖ associated with the actual data stored in a file (often calculated as a numerical function of the values within the bits of the files) that can be used to verify that the data stored corresponds to the data shared or received. When the checksum calculated for a retrieved block does not equal the stored checksum of that block, it is considered an integrity error. In that case, the requested block will need to be retrieved from a replica instead.
• Metadata replication: The metadata files are also subject to failure, and HDFS can be configured to maintain replicas of the corresponding metadata files to protect against corruption.
• Snapshots: This is incremental copying of data to establish a point in time to which the system can be rolled back.
These concepts map to specific internal protocols and services that HDFS uses to enable a large-scale data management file system that can run on commodity hardware components. The ability to use HDFS solely as a means for creating a scalable and expandable file system for maintaining rapid access to large datasets provides a reasonable value proposition from an Information Technology perspective:
• decreasing the cost of specialty large-scale storage systems;
• providing the ability to rely on commodity components;
• enabling the ability to deploy using cloud-based services;
• reducing system management costs.

MAPREDUCE AND YARN

Here the general concept of job control and management are introduced. In Hadoop, MapReduce originally combined both job management and oversight and the programming model for execution. The MapReduce execution environment employs a master/slave execution model, in which one master node (called the JobTracker) manages a pool of slave computing resources (called TaskTrackers) that are called upon to do the actual work.
The role of the JobTracker is to manage the resources with some specific responsibilities, including managing the TaskTrackers, continually monitoring their accessibility and availability, and the different aspects of job management that include scheduling tasks, tracking the progress of assigned tasks, reacting to identified failures, and ensuring fault tolerance of the execution. The role of the TaskTracker is much simpler: wait for a task assignment, initiate and execute the requested task, and provide status back to the JobTracker on a periodic basis.

Different clients can make requests from the JobTracker, which becomes the sole arbitrator for allocation of resources.

There are limitations within this existing MapReduce model. First, the programming paradigm is nicely suited to applications where there is locality between the processing and the data, but applications that demand data movement will rapidly become bogged down by network latency issues. Second, not all applications are easily mapped to the MapReduce model, yet applications developed using alternative programming methods would still need the MapReduce system for job management. Third, the allocation of processing nodes within the cluster is fixed through allocation of certain nodes as ―map slots‖ versus ―reduce slots.‖ When the computation is weighted toward one of the phases, the nodes assigned to the other phase are largely unused,
resulting in processor underutilization.

This is being addressed in future versions of Hadoop through the segregation of duties within a revision called YARN. In this approach, overall resource management has been centralized while management of resources at each node is now performed by a local NodeManager.

In addition, there is the concept of an ApplicationMaster that is associated with each application that directly negotiates with the central ResourceManager for resources while taking over the responsibility for monitoring progress and tracking status. Pushing this responsibility to
the application environment allows greater flexibility in the assignment of resources as well as be more effective in scheduling to improve node utilization.

Last, the YARN approach allows applications to be better aware of the data allocation across the topology of the resources within a cluster.

This awareness allows for improved colocation of compute and data resources, reducing data motion, and consequently, reducing delays associated with data access latencies. The result should be increased scalability and performance.

THE MAPREDUCE PROGRAMMING MODEL

We can use the Hadoop MapReduce programming model as an example. One can read more about MapReduce at Apache‘s MapReduce Tutorial Page.1 Note that MapReduce, which can be used
to develop applications to read, analyze, transform, and share massive amounts of data is not a database system but rather is a programming model introduced and described by Google researchers for parallel, distributed computation involving massive datasets (ranging from hundreds of terabytes to petabytes).

Application development in MapReduce is a combination of the familiar procedural/imperative approaches used by Java or C11 programmers embedded within what is effectively a functional language programming model such as the one used within languages like Lisp and APL. The similarity is based on MapReduce‘s dependence on two basic operations that are applied to sets or lists of data value pairs:

1. Map, which describes the computation or analysis applied to a set of input key/value pairs to produce a set of intermediate key/value pairs.

1. Reduce, in which the set of values associated with the intermediate key/value pairs output by the Map operation are combined to provide the results.

A MapReduce application is envisioned as a series of basic operations applied in a sequence to small sets of many (millions, billions, or even more) data items. These data items are logically organized in a way that enables the MapReduce execution model to allocate tasks that can be executed in parallel. The data items are indexed using a defined key into ,key, value. pairs, in which the key represents some grouping criterion associated with a computed value. With some

applications applied to massive datasets, the theory is that the computations applied during the Map phase to each input key/value pair are independent from one another. Figure 8.1 shows how Map and Reduce work.

Combining both data and computational independence means that both the data and the computations can be distributed across multiple storage and processing units and automatically parallelized. This parallelizability allows the programmer to exploit scalable massively parallel processing resources for increased processing speed and performance.

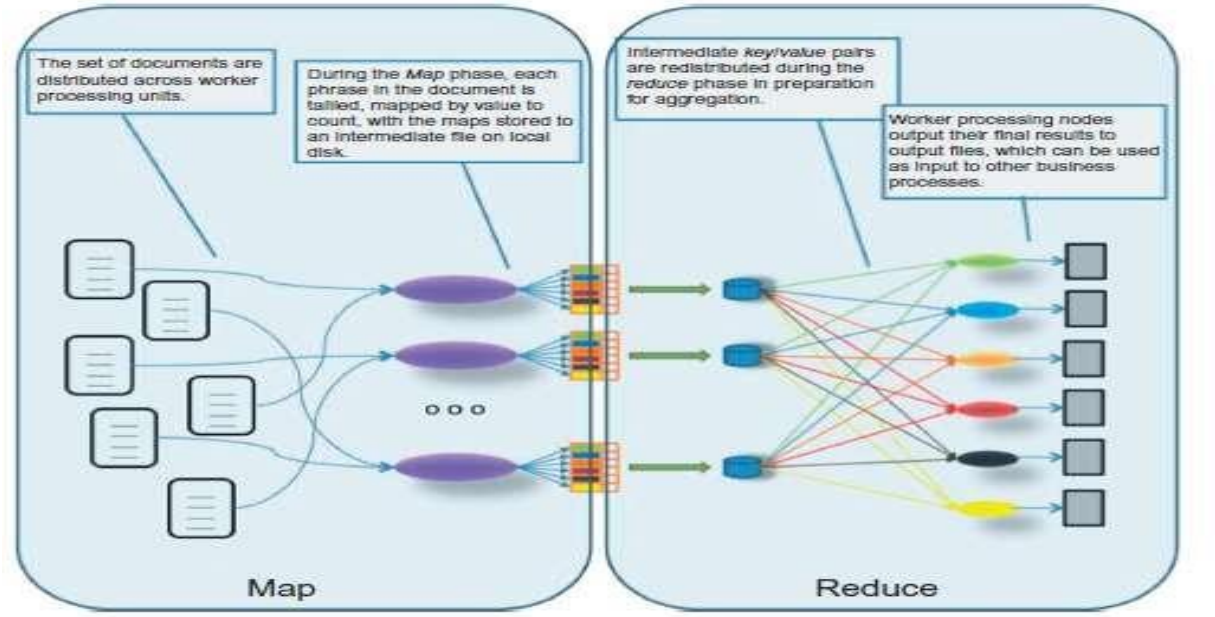


Figure 8.1 How Map and Reduce work.