Leveraging Hadoop-Based Big Data Architectures for a Scalable, High-Performance Analytics Platform

Mrs. B. Manimekala
Assistant Professor, Department of Computer Science,
AJK College of Arts and Science,
Coimbatore

Abstract:- **A general overview of traditional Hadoop architectures designed to deliver high-performance and scalable big data analytics. It is intended to provide a basis of understanding for interested data center architects and as a starting point for a deeper implementation engagement. The goal is to empower IT to enable an infrastructure that can provide immediate and deep business intelligence for decision making and agility. In addition, the increased use of sensors for everything from traffic patterns, purchasing behaviors, and real-time inventory management is a primary example of the massive increase in data. Much of this data is gathered in real time and provides a unique and powerful opportunity if it can be analyzed and acted upon quickly. Machine-to-machine interchange is another often unrecognized source of big data. The rise of security information management (SIM) and the Security Information and Event Management (SIEM) industry is at the heart of gathering, analyzing, and proactively responding to event data from active machine log files. At the heart of this trend is the ability to capture, analyze, and respond to data and data trends in real time.**

Keywords: Big Data, Hadoop, SIM, SIEM

WHEN DOES DATA GET BIG?

People, devices and things are constantly generating massive volumes of data. At work people create data, as do children at home, students at school, people and things on the move, as well as objects that are stationary. Devices and sensors attached to millions of things take measurements from their surroundings, providing up-to-date readings over the entire globe – data to be stored for later use by countless different applications.

WHAT IS BIG DATA?

Big data refers to the collection and subsequent analysis of any significantly large collection of data that may contain hidden insights or intelligence (user data, sensor data, and machine data). When analyzed properly, big data can deliver new business insights, open new markets, and create competitive advantages.

Big-data technologies are a new generation of methods and architectures designed to extract value from masses of different data types through high-velocity capture, discovery and analysis. To analyze the large volumes of bytes associated with big data in a cost-efficient manner, a shift in the common approach to computer architecture is needed. By moving from costly hardware to commodity computing, operators will be able to meet the cost requirements for massive amounts of data storage and heavy server-processing power.

THREE V'S OF BIG DATA

Volume – big data comes in one size: XXL. Enterprises and operators are saturated with data; they amass huge

Fig No.1 Big Data Dimensions – Three V's

amounts of it daily, and available storage cannot handle these volumes.

Velocity – data needs to be used quickly to maximize business benefit before the value of the information is lost.

Variability – data can be structured, unstructured, semi-structured or a mix of all three. It comes in many forms including text, audio, video, click streams and log files. Some of it is new and some is old.

TYPES OF DATA

The value that can be derived from using big-data technologies depends on the use case and the data associated with it. Apart from volume and velocity, the value that can be gained from the variability of data tends to be overlooked. The structured the data, the greater the requirement to apply big-data technologies. Variability is typically categorized into three different data types:

Structured – data is well organized, there are several choices for abstract data types, and references such as relations, links and pointers are identifiable.
Unstructured – data may be incomplete and/or heterogeneous, and often originates from multiple sources. It is not organized in an identifiable way, and typically includes bitmap images or objects, text and other data types that are not part of a database.
Semi-structured – some data is organized, containing tags or other markers to separate semantic elements, but unstructured data may also be present.

A SCALABLE DATA INFRASTRUCTURE

Another unique characteristic of big data is that, unlike large data sets that have historically been stored and analyzed, often through data warehousing, big data is made up of discretely small, incremental data elements with real-time additions or modifications. It does not work well in traditional, online transaction processing (OLTP) data stores or with traditional SQL analysis tools. Big data requires a flat, horizontally scalable database, often with unique query tools that work in real time with actual data.

Components	Traditional Data	Big Data
Architecture	Centralized	Distributed
Data Volume	Terabytes	Petabytes to exabytes
Datatype	Structured or Transactional	Unstructured or Semi-structured
Data Relationships	Known relationships	Complex Unknown Relationships
Data Model	Fixed Schema	Schema-less

Table 1: Big Data Vs Traditional Datatypes

INSIDE THE TECHNOLOGY

Big-data technologies are usually engineered from the bottom up with two things in mind: scale and availability. Consequently, most solutions are distributed in nature and introduce new programming models for working with large volumes of data. Because most of the legacy database models cannot be effectively used for big data, the current approach to ensuring availability and partitioning needs to be revised.

NOSQL

Like key-value storage of semi-structured data, NoSQL systems are designed with specific characteristics in mind, such as relaxed models of consistency. They run applications that tend to be read/write-intensive.

To take advantage of scaling capacity as new nodes are added to a network, many NoSQL databases are designed to expand horizontally and run on low-cost commodity hardware. NoSQL databases have far more relaxed or even nonexistent data-model restrictions. Such databases allow applications to store almost any kind of structure in a data element, and the responsibility for maintaining the logical data structure is transferred to the application.

Most NoSQL databases are key-value stores that hold schema-less collections of entities that do not necessarily share the same properties. The data consists of a string containing the key, and the actual data is considered to be the value in the key-value relationship.

CAP TRIANGLE

Consistency – all nodes see the same data at the same time.
Availability – every operation results in a response (success or failure).
Partition tolerance – the system continues to operate when individual components are unavailable, or when messages are lost.

Big Data

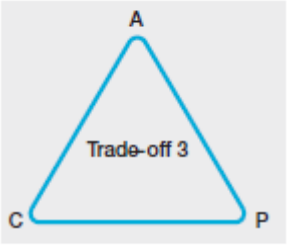

Fig No. 2 Cap Triangle

Technologies (Hadoop)

The driving force behind an implementation of big data is the software—both infrastructure and analytics. Primary in the infrastructure is Hadoop. Hadoop is the big data management software infrastructure used to distribute, catalog, manage, and query data across multiple, horizontally scaled server nodes. Yahoo! created it based on an open source implementation of the data query infrastructure (originated at Google) called MapReduce. It has a number of commercially supported distributions from companies such as MapR Technologies and Cloudera.

Hadoop is a framework for processing, storing, and analyzing massive amounts of distributed unstructured data. As a distributed file storage subsystem, Hadoop Distributed File System (HDFS) was designed to handle petabytes and exabytes of
data distributed over multiple nodes in parallel.

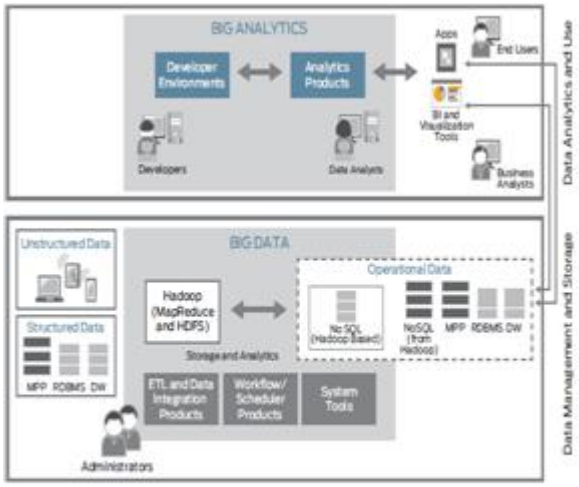

Fig No.3 High Level Structure of Hadoop Deployment

WHAT DOES BIG DATA MEAN TO IT?

As an IT solution, big data mirrors the growth in content and data source, as well as the pervasiveness of technology in our everyday lives. As more and more of what we do is both connected to and often empowered by a network—and the devices that we connect to are themselves powered by an array of sensors - should expect that the ongoing stream of data will grow. Within data centers, every node (servers, storage, and applications) generates a tremendous number of log files and isolated data streams that also can be collected, collated, and analyzed. With storage costs dropping, the cost associated with saving and leveraging even the most mundane data becomes a nonissue.

THE HADOOP CLUSTER

Hadoop, which includes a distributed file system known as Hadoop Distributed File System (HDFS) and MapReduce, is a critical big data technology that provides a scalable file system infrastructure and allows for the horizontal scale of data for quick query, access, and data management. At its most basic level, a Hadoop implementation creates four unique node types for cataloging, tracking, and managing data throughout the infrastructure: data node, client node, name node, and job tracker. The capabilities of these four types are generally as follows:

Data node - The data nodes are the repositories for the data, and consist of multiple smaller database infrastructures that are horizontally scaled across compute and storage resources through the infrastructure. Larger big data repositories will have numerous data nodes. The critical architectural concern is that unlike traditional database infrastructure, these data nodes have no necessary requirement for locality to clients, analytics, or other business intelligence.

Client - The client represents the user interface to the big data implementation and query engine. The client could be a server or PC with a traditional user interface.

Name node - The name node is the equivalent of the address router for the big data implementation. This node maintains the index and location of every data node.

Job tracker - The job tracker represents the software job tracking mechanism to distribute and aggregate search queries across multiple nodes for ultimate client analysis.

Within each data node, there may exist several tens of server or data storage elements and its own switching tier connecting each storage element with the overall Hadoop cluster. Big data infrastructure purposely breaks the data into horizontally scaled nodes, which naturally adds latencies across nodes. This is important as locality and network hops represent potential latencies in the architecture.

WHY IS THIS IMPORTANT?

Distributed data of today's big data and cloud architectures places a tremendous burden on connecting nodes rather than connecting clients. For every one-client interaction, there may be hundreds or thousands of server and data node interactions. This is counter to the original client/server network architectures built over the last 20 years. Those architectures assumed that the client, rather than the backend infrastructure, supplied much of the computational overhead. With that in mind, network administrators should consider distributed systems such as Hadoop.

Hadoop and most cloud computing infrastructures today run as a cluster and handle huge data sets, which are distributed over multiple nodes. Hadoop clusters run tasks in parallel and scale-out fashion. Although Hadoop is agnostic to network infrastructure, it places the following requirements on the network:

Data locality - The data shuffle-and-sort operation between the distributed Hadoop nodes running parallel jobs causes east-west network traffic that can be adversely affected by suboptimal network connectivity. The network has to provide high bandwidth, low latency, and any-to-any connectivity between the nodes for optimal Hadoop performance.

Scale-out - Deployments might start with a small cluster and then scale out over time as the customer realizes initial success and then needs change. The underlying network architecture also should scale seamlessly with Hadoop clusters and provide predictable performance.

Increased east-west traffic - As previously described, traffic patterns range from 1-to-1, 1-to-many, many-to-1, and many-to-many. These flows are demanded by a combination of unicast/multicast flows between multiple Hadoop nodes that run in parallel. This requires a high bandwidth, low latency network infrastructure for efficient communication between Hadoop nodes.

As a result, it is critical in the network architecture to prioritize locality, high-performance horizontal scalability, and direct server node to server node connectivity. In addition, with the latencies prevalent in tiered node hops, it is evident that a new network architecture is required for high-performance scale.

MAP REDUCE

Another significant aspect of big-data technologies is MapReduce programming. By splitting a problem into many smaller ones that can be processed simultaneously on a number of nodes, MapReduce offers an extremely efficient solution when the database application needs to aggregate all occurrences of a word or phrase in a very large document database, for example.

MapReduce is suitable for solving problems that can be broken down into smaller independent sub-problems with results that can be aggregated to produce a single answer. Google owns the patent for this programming method, which is based on the map and reduces functions commonly used in parallel programming.

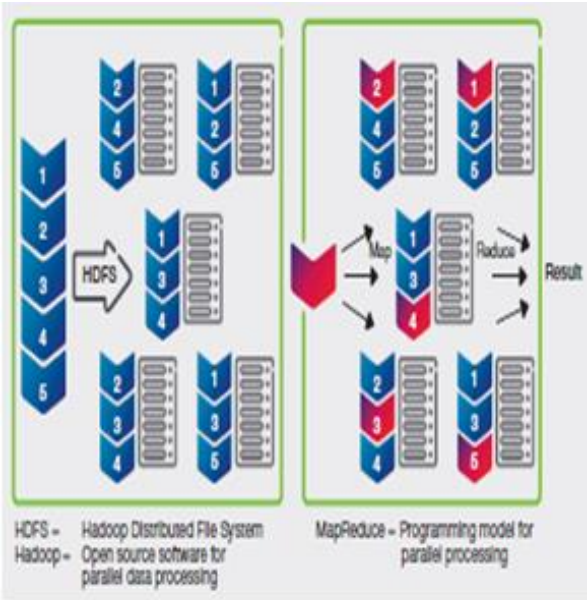

Fig No.4 A Parallel Data – Processing System

A Scalable Switching Fabric Network Infrastructure

One of the biggest evolutions in the networking industry in the last few years is the introduction of point-to-point switching fabric network solutions. The best way to describe a network switching fabric and its benefits is to refer to
Wikipedia's definition:
"Switched fabric, switching fabric, or just fabric, is a network topology where network nodes connect with each other via one or more network switches (particularly via crossbar switches, hence the name).
The term is popular in telecommunication, Fibre Channel storage area networks, and other high-speed
networks. The term is in contrast to a broadcast medium such as early forms of Ethernet. Switched fabrics can offer better total throughput than broadcast networks because traffic is spread across multiple physical links.
"A switched fabric should be able to function as a simple point-to-point interconnect and also scale to handle thousands of nodes. The term fabric derives its name from its topological representation. As the
data paths between the nodes of a fabric are drawn out, the lines cross so densely that the topology map
is analogous to a cloth.
"The advantage of the switched fabric is typically one of overall system bandwidth and performance
versus connectivity between individual devices."

Understanding the benefits of a data center fabric is critical in creating the highest performance/lowest latency connections between big data nodes. The benefit of switch fabric infrastructure compared to more traditional tree architecture is obvious and startling. First, since fabric architectures create a point-to-point connection between nodes
with a single hop in the switching infrastructure, inherently this architecture will significantly reduce latencies between nodes. Second, for those looking to provide a seamless high-performance interconnect with the easiest switching management, virtualizing the switching fabric is also an essential feature, as it allows multiple networking components to behave as a single component.

Flat network scalability also is a major benefit of implementing a data center fabric. The inherent virtual domains and point-to-point connections allow a company to seamlessly merge new data sources into the cluster without rewiring the entire data center. This benefit significantly eases the expansion of the system and allows for rapid enhancement of the analytic environment.

BIG DATA USE CASES

There are many examples of big data use cases in virtually every industry imaginable. Some businesses have been more receptive of the technologies and faster to integrate big data analytics into their everyday business than others. It is evident that organizations embracing this technology not only will see significant first-mover advantages but will be
considerably more agile and cutting edge in the solutions and adaptability of their offerings. Use case examples of big data solutions include:

Financial services providers are adopting big data analytics infrastructure to improve their analysis of customers to help determine eligibility for equity capital, insurance, mortgage, or credit.

Airlines and trucking companies are using big data to track fuel consumption and traffic patterns across their fleets in real time to improve efficiencies and save costs.

Healthcare providers are managing and sharing patient electronic health records from multiple sources—imagery, treatments, and demographics—and across multiple practitioners. In addition, pharmaceutical companies and regulatory agencies are creating big data solutions to track drug efficacy and provide more efficient and shorter drug development processes.

Telecommunications and utilities are using big data solutions to analyze user behaviors and demand patterns for a better and more efficient power grid. They are also storing and analyzing environmental sensor data to provide

insight into infrastructure weaknesses and provide better risk management intelligence.

Media and entertainment companies are utilizing big data infrastructure to assist with decision making around customer lifecycle retention and predictive analysis of their user base, and to provide more focused marketing and customer analytics.

CONCLUSION

The data available to operators through their networks presents them with an opportunity and a business-intelligence edge over other players. As is often the case, with opportunity comes challenge, and for big data this challenge comprises the volume and diversity of the data – and the fact that both are expected to grow substantially in the next few years. The value of the information in the data is significant, but the costs involved in obtaining it using current technology are inhibitive.

Consequently, big-data technology is an important part of the puzzle for operators wanting to leverage value from the large volumes of data in their possession in a cost-efficient way. Applying big-data technologies has the side effect of transferring some complexity from the database to the application.

REFERENCES

1. John Gantz and David Reinsel, IDC, June 2011, Extracting Value from Chaos, available at: http:// www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
2. Ericsson, June 2012, Traffic and Market Report, available at: http://www.ericsson.com/ res/docs/2012/traffic_and_market_report_june_2012.pdf
3. IBM, 2012, What is Big Data? available at: http:// www-01.ibm.com/software/data/bigdata/
4. Ericsson, The Networked Society site: http://www.ericsson.com/networkedsociety.
5. Ericsson, 2011, The Networked Society City Index, available at: http://www.ericsson.com/ networkedsociety/lab/research/city-index/
6. Dan Pritchett, Association for Computing Machinery (ACM), 2008, BASE – an ACID Alternative, available at: http://queue.acm.org/detail.cfm?id=1394128
7. Rasmus Päivärinta, Aalto University, Finland, Applicability of NoSQL databases to Mobile Networks, available at: http://www.cse.hut.fi/fi/opinnot/T-110.5121/2011/lisatty-files/nosql.pdf
8. Mona Matti and Tor Kvernvik, Applying Big Data network Architecture, at Ericsson Review, 284 23-3181 | Uen
9. White Paper - Introduction to Big Data: Infrastructure and Networking Considerations

AUTHOR

She completed her MCA M.Phil. Presently she was working as an Assistant Professor in Department of Computer Science, AJK College of Arts and Science. She has eight years of teaching experience. She presented three papers in National Conference and two papers in International Journals.