ISSN:0975-9646

Intel Distribution for Apache Hadoop

Tarush Jain, Rohan Somni

IT Department, Pune University
Pune, India

Abstract: **Big Data and its implications is a current hot topic in the overall global scenario. In this paper we describe big data, as well as Apache Hadoop the most widely used framework for processing big data, which is at the fore of cutting edge technology-and for which new implementations are still being devised. We introduce Intel Distribution for Apache Hadoop as one such implementation. An overview of its advantages and uses, and how it manages to improve drastically on Hadoop's big data processing performance using specific encryption techniques and specialized Intel processors is given.**

I. INTRODUCTION:

With continuous technological advances and breakthroughs, data has obtained paramount importance in all fields. Now cheaper than ever, managing and processing it effectively is the main objective of organizations in all fields. In light of these developments, 'big data' is introduced and discussed. As a simultaneous supplement and solution to that discussion Apache Hadoop is explained, along with how it addresses the issues posed by big data.

A systematic study of the technical aspects of Hadoop architecture, paradigms used, its File System and internal mechanism is done.

II. BIG DATA

The current age is the Age of Technology. Data is now cheap, and the amount of data available is huge. A tremendous data explosion has occurred over the last decade and as a result the total amount of data currently present roughly equals 1.2 zeta bytes, and this amount will continue to grow at an exponential rate.

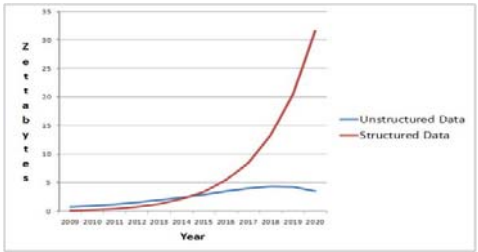

Fig.1 Growth of data

The majority of this data is unstructured application data, which cannot be easily processed using databases. Healthcare, transportation, finance, energy and resource conservation, environmental sustainability, and homeland security are but a few of society's grand challenges that look to information systems for efficient management and, more importantly, quality outcomes and solutions- along with other businesses and commercial organizations.

Regardless of the specific challenge, underlying technologies and evolving user requirements continue to expand both data volume and variety. Data is coming from every imaginable source, often in real time, and the stakeholder demand for quality outcomes has never been higher.

This is where the concept of big data comes into play. Big data can be defined as:

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- Big is a relative term dependent on the individual or organization. Big data is when the normal application of current technology doesn't enable users to obtain timely, cost-effective, and quality answers to data-driven questions.

In layman terms, the absurdly large amount of data which is big data, cannot be handled using current means with sufficient efficiency to meet the required objectives of the organization in question.

Big data can be considered to be a combination of big throughput and big analytics.

The former includes the problems associated with storing and manipulating large amounts of data (in relation to the available resources, as mentioned) and the latter those concerned with transforming this data into knowledge. The information that constitutes big data can be represented by the 3 V's: Variety, Velocity, and Volume.

- **Variety:** Data today comes in all types of formats – from traditional databases to hierarchical data stores created by end users and OLAP systems, to text documents, email, meter-collected data, video, audio, stock ticker data and financial transactions. By some estimates, 80 percent of an organization's data is non-numeric. But it still must be included in analyses and decision making.
- **Velocity:** According to Gartner, velocity "means both how fast data is being produced and how fast the data must be processed to meet demand." RFID tags and smart metering are driving an increasing need to deal with torrents of data in near-real time. Reacting quickly enough to deal with velocity is a challenge to most organizations.
- **Volume:** Many factors contribute to the increase in data volume – transaction-based data stored through the years, text data constantly streaming in from

social media, increasing amounts of sensor data being collected, etc. In the past, excessive data volume created a storage issue. But with today's decreasing storage costs, other issues emerge, including how to determine relevance amidst the large volumes of data and how to create value from data that is relevant.

III. APACHE HADOOP

Hadoop, also known as Apache Hadoop, is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware. The Hadoop framework transparently provides both reliability and data motion to applications. Hadoop is written in the Java programming language and is a top-level Apache project being built and used by a global community of contributors. Hadoop enables applications to work with thousands of computation-independent computers and petabytes of data. It was derived from Google's MapReduce and Google File System (GFS) papers.

Hadoop implements a computational paradigm named map/reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both map/reduce and the distributed file system are designed so that node failures are automatically handled by the framework. Hadoop implementation is generally done with clusters, which are loosely connected machines sharing no hardware or memory segments. in any sophisticated system, there is a mixture of structured and unstructured data which is hard to fit into a single database. Hadoop is designed to run on a large number of machines that don't share any memory or disks. It spreads data over the organization's servers by breaking it into pieces and keeps track of where it is stored.

Hadoop consists of the Hadoop Common which provides access to the filesystems supported by Hadoop. The Hadoop Common package contains the necessary JAR files and scripts needed to start Hadoop. The package also provides source code and documentation. A small Hadoop cluster will include a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode, and DataNode. A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes, and compute-only worker nodes; these are normally only used in non-standard applications. Hadoop requires JRE 1.6 or higher. The standard startup and shutdown scripts require ssh to be set up between nodes in the cluster.

In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the file system index, and a secondary NameNode that can generate snapshots of the name node's memory structures, thus preventing filesystem corruption and reducing loss of data.

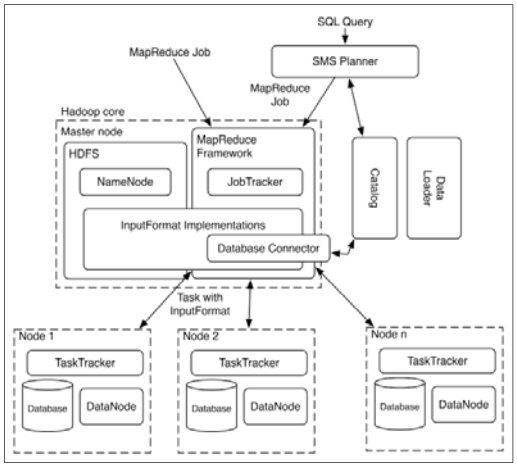

Fig.2 Structure of Hadoop

In organizations, thousands of servers both host directly attached storage and execute user applications. The Hadoop cluster needs to be set according to the processing required. The main workflow is as follows:

- Load data into the cluster
- Analyze the data (Map Reduce)
- Store results into the cluster (HDFS writes)
- Read results from the cluster (HDFS reads)

The Job Tracker must be assigned the instructions regarding the processing required in the form of Java code, which it then forwards to the corresponding nodes that contain the relevant data. Setting a Hadoop cluster requires the operator to strike a balance between encryption and performance, according to the necessity of either for data processing. Intel® Distribution for Hadoop provides improved performance on this front, which is discussed later.

IV. ADVANCED ENCRYPTION STANDARD (AES)

AES (Advanced Encryption Standard) is an encryption standard adopted by the U.S. government starting in 2001. It is widely used across the software ecosystem to protect network traffic, personal data, and corporate IT infrastructure. AES is a symmetric block cipher that encrypts/decrypts data through several rounds. The new 2010 Intel® Core™ processor family (code name Westmere) includes a set of new instructions, **Intel® Advanced Encryption Standard (AES) New Instructions (AES-NI)**. The instructions were designed to implement some of the complex and performance intensive steps of the AES algorithm using hardware and thus accelerating the execution of the AES algorithms. AES-NI can be used to accelerate the performance of an implementation of AES by 3 to 10x over a completely software implementation.

The AES algorithm works by encrypting a fixed block size of 128 bits of plain text in several rounds to produce the final encrypted cipher text. The number of rounds (10, 12, or 14)

used depends on the key length (128b, 192b, or 256b). Each round performs a sequence of steps on the input state, which is then fed into the following round. Each round is encrypted using a sub key that is generated using a key schedule. This leads to improved performance as well as security, and is implemented in Intel's Distribution for Apache Hadoop.

V. INTEL DISTRIBUTION FOR APACHE HADOOP

(Intel® Distribution) is a software platform that provides distributed processing and data management for enterprise applications that analyse massive amounts of diverse data. Intel Distribution is an open source software product that includes Apache Hadoop and other software components along with enhancements and fixes from Intel. Proven in production at some of the most demanding enterprise deployments in the world, Intel Distribution is supported by a worldwide engineering team with access to expertise in the entire software stack as well as the underlying processor, storage, and networking components.

Key Features:

- Up to 30x boost in Hadoop performance with optimizations for Intel® Xeon processors, Intel® SSD storage, and Intel® 10GbE networking
- Data confidentiality without a performance penalty with encryption and decryption in HDFS enhanced by Intel® AES-NI and role-based access control with cell-level granularity in HBase
- Multi-site scalability and adaptive data replication in HBase and HDFS
- Up to 3.5x improvement in Hive query performance
- Support for statistical analysis with R connector
- Enables graph analytics with Intel® Graph Builder
- Enterprise-grade support and services from Intel

Encryption can be done as required without sacrificing processing speed and cluster performance.

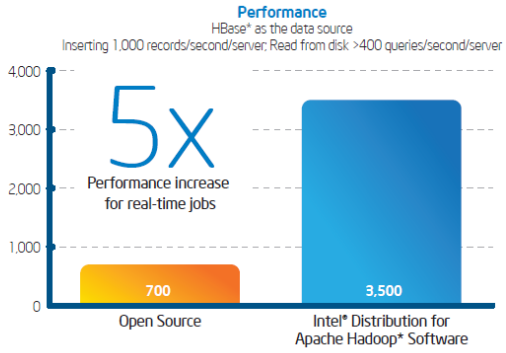

Fig.3 Performance measures of Intel® Distribution

This solution has been provided primarily to assist the healthcare industries which have to deal with big data, on a regular basis, but also cannot afford to neglect data encryption. These enhancements support the velocity at which researchers, providers, and payors can analyze data and make informed decisions based on that data. For example, emergency medical personnel who can rapidly access a patient's EHR for allergy information can make better care decisions for that patient. Easy access to cross-regional—or even international—public health records can help public health officials track and respond to disease outbreaks.

VI. CONCLUSION

In this age, data is money. Growth of data is exponential, and will continue to increase as technology develops. Hadoop is the currently the primary tool for processing and handling big data, and many developments are being made to improve its efficiency. It offers a powerful tool for analysing large and diverse data sets, yet the lack of integrated support for strong data security has been a serious roadblock to implementation for many businesses. The Intel Distribution for Apache Hadoop software provides an answer: a comprehensive, enterprise-ready software platform for big data analytics that is highly optimized for performance, stability, and manageability.

The scope of cutting edge technologies for big data analytics and processing will continue to grow, as new innovative applications are being developed to provide ever increasing efficiency and will be the key in future technological developments to come. We have attempted to provide an insight into the functionality and architecture of Apache Hadoop and outline Intel Distribution to emphasize the advantages that continued development in this field will bring.

REFERENCES:

[1] White Paper, "Intel® Distribution for apache hadoop* software helps cure big data woes", Healthcare IT.
[2] Solution brief, "Intel ® Encryption for Hadoop" .
[3] Product Brief, "Intel ® Distribution for Hadoop Software".
[4] Bhandarkar, M., "MapReduce programming with apache Hadoop", Parallel & Distributed Processing (IPDPS), 2010 IEEE
[5] Cohen, J.C. ; Acharya, S., "Incorporating hardware trust mechanisms in Apache Hadoop: To improve the integrity and confidentiality of data in a distributed Apache Hadoop file system: An information technology infrastructure and software approach", Globecom Workshops (GC Wkshps), 2012 IEEE
[6] Mukherjee, Anirban ; Datta, Joydip ; Jorapur, Raghavendra ; Singhvi, Ravi ; Haloi, Saurav ; Akram, Wasim, "Shared disk big data analytics with Apache Hadoop", High Performance Computing (HiPC), 2012 19th International Conference.
[7] Kousiouris, G. ; Vafiadis, G. ; Varvarigou, T., "A Front-end, Hadoop-based Data Management Service for Efficient Federated Clouds", Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference.
[8] Chih-Chung Lu ; Shau-Yin Tseng, "Integrated design of AES (Advanced Encryption Standard) encrypter and decrypter", Application-Specific Systems, Architectures and Processors, 2002.
[9] White Paper, "Intel® Advanced Encryption Standards(AES) New Instructions Set.