(intel®)

Intel® Distribution for Apache Hadoop* Software

Open Platform for Next-Gen Analytics

Intel® Distribution for Apache Hadoop* software (Intel® Distribution) is a software platform that provides distributed data processing and data management for enterprise applications that analyze massive amounts of diverse data. The Intel Distribution includes Apache Hadoop and other software components with enhancements from Intel. Proven in production at some of the most demanding enterprise deployments in the world, the Intel Distribution is supported by experts at Intel with deep optimization experience in the Apache Hadoop software stack as well knowledge of the underlying processor, storage, and networking components.
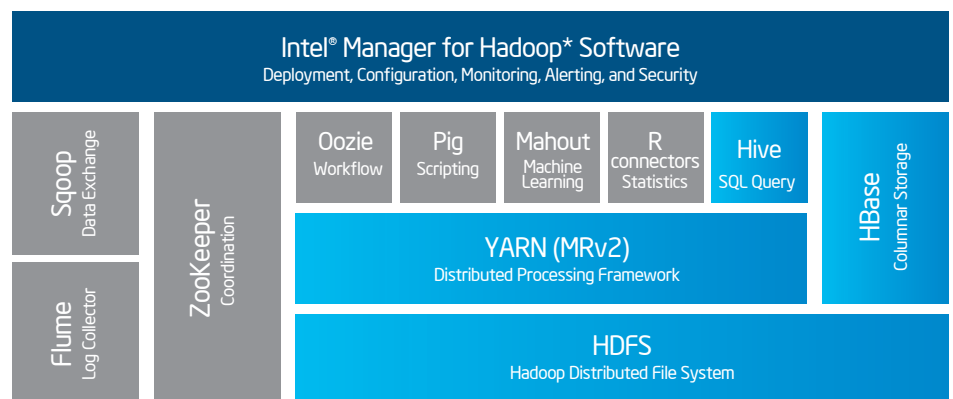

Figure 1. Intel Distribution for Apache Hadoop software components.

Why Intel Software for Apache Hadoop?

We believe that every organization and individual should have the ability to generate value from all the data they can access. The Intel Distribution for Apache Hadoop is designed to enable the widest range of use cases on Hadoop by delivering the performance and security that enterprises need. Along with the Intel® Xeon processor, SSD, and Intel® 10GbE networking, the Intel Distribution offers a robust platform upon which the ecosystem can innovate in delivering new analytics solutions. Intel delivers platform innovation in open source and is committed to supporting the Apache developer community with code and collaboration.

Keep Pace with Open Innovation

Apache Hadoop continues to evolve into a powerful platform for data processing and data management for a number of analytics use cases from detecting credit fraud, serving targeted ads and online product placement, data mining genome sequences, improving the efficiency of power grids, to modeling traffic patterns.

As businesses, government agencies, and other data-driven organizations use Apache Hadoop for a growing number of applications, they need an open platform that not only keeps pace with not only the growth in data but also the technological advancements in processor, memory, storage, networking, and fabric that are occurring in parallel.

Intel Distribution for Apache Hadoop is designed to reflect ongoing innovation in the hardware platform by delivering that value in the Apache Hadoop software stack. Software engineers at Intel continue to enable advanced hardware capabilities in every layer of the software stack – from the hypervisor and Linux operating system to Java, Hadoop, HDFS, HBase, and Hive. This robust platform enables the entire software ecosystem to build innovative solutions for analytics by delivering unique value where it matters to their customers. By delivering hardware-enhanced software value in the open at time to market with the hardware capabilities, the Intel Distribution accelerates innovation in the entire ecosystem around Apache Hadoop.

Run High Performance Analytics on a Balanced Platform

Apache Hadoop is designed and optimized for commonly available hardware. Using the latest generation of hardware components comprising the systems' compute, memory, storage, and networking resources delivers significant gains in performance.

- **Processor:** The Intel® Xeon® processor E5-1600/2600 product families deliver up to 50 percent higher performance than the previous generation.

- **Storage:** Performance improvements in storage I/O are ongoing, and the costs to gain the performance advantages of solid-state drives (SSD) keep dropping. The use of SSD with Intel® Cache Acceleration Software can reduce the processing time by 80% for some workloads.

- **Network:** 10 Gigabit Ethernet (10GbE) networking demonstrates its value in the form of high levels of network utilization in the Hadoop cluster. The full use of greater bandwidth can reduce time to ingest and to export data by 80 percent. Moreover, the cost per gigabit of bandwidth with 10GbE is now much lower than 1GbE, making it a natural choice for big data.
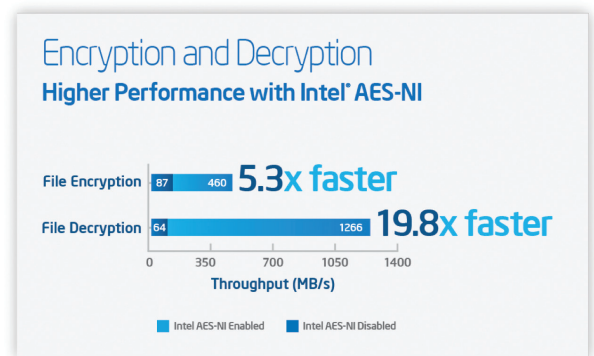
Much of the performance gain from the underlying hardware requires deep optimization in the software as well as careful tuning of Hadoop configuration parameters. The Intel Distribution is optimized with the latest Intel® processor, storage, and networking hardware components to ensure that the platform delivers balanced performance for the widest range of use cases.

Secure Data without Compromising Performance

Apache Hadoop offers a powerful tool for analyzing 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. Encryption and decryption are compute-intensive processes that traditionally add considerable latency and consume substantial processing resources.

The Intel Distribution helps to eliminate much of the latency and greatly reduce the load on the processors using Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI), a technology that is built into Intel® Xeon® processors. Intel AES-NI provides seven instructions that accelerate the most complex and compute-intensive steps of the AES algorithms. It also helps to make encryption stronger by protecting against "side channel" snooping attacks to break encryption codes.

The Intel Distribution for Apache Hadoop accelerates data encryption by up to 6.2x and data decryption by up to 19.8x. Using a version of OpenSSL* optimized by Intel to take advantage of Intel AES-NI, encryption is transparent to users, can be applied on a file-by-file basis, and works in combination with external key management applications.

Encryption and Decryption
Higher Performance with Intel® AES-NI

File Encryption — 87 / 460 — **5.3x faster**
File Decryption — 64 / 1266 — **19.8x faster**

Throughput (MB/s): 0, 350, 700, 1050, 1400

■ Intel AES-NI Enabled   ■ Intel AES-NI Disabled

Security-sensitive files are encrypted by external applications before they arrive at the Hadoop cluster and are loaded into HDFS. Each file arrives with the corresponding encryption key. When an encrypted file enters the Apache Hadoop environment, it remains encrypted in HDFS. It is then decrypted as needed for processing and re-encrypted before it is moved back into storage. The results of the analysis, including intermediate results, are also encrypted. Data and results are always encrypted, while in transit and at rest.

Enterprises and government agencies can now store and process encrypted data on Apache Hadoop and HDFS without having to trade off performance by deploying the Intel Distribution.

Proven in Data-Intensive Enterprise Environments

Tested in some of the most rigorous lab environments and deployed in production at enterprise scale, the Intel Distribution can meet the most demanding requirements of an enterprise. Intel has been actively working with the Hadoop framework for several years. We have worked closely with some the largest telecommunications, e-commerce, and government agencies in China to develop and deploy solutions over the last year. In the US, the Intel Distribution is deployed at a number of sites in production and in pilot as the platform for big data applications ranging from genomic analysis to information management for financial services.

For example, NextBio, a software company that provides a cloud service for drug companies, is using the Intel Distribution to mine billions of data points in real-time. This allows the researchers to achieve advancements in medicine by more quickly analyzing molecular data that could result in new pharmaceuticals or reduce the time it takes to conduct clinical trials.

Maximize Value with Subscription and Support

Take advantage of the experience and expertise of the engineers standing behind the Intel Distribution for Apache Hadoop to keep your deployment running smoothly. When you purchase a subscription for the Intel Distribution, you get updates to the software and access to telephone support. We see ourselves as a trusted advisor, and our first priority is to get you the information you need, whether its issue resolution or best practices.

Intel provides two tiers of support for Intel Distribution for Apache Hadoop and Intel Manager included in the subscription.

Subscription and Support		
FEATURES	**PREMIUM**	**STANDARD**
NUMBER OF SUPPORT INCIDENTS	Unlimited	Unlimited
NUMBER OF SUPPORT CONTACTS	Unlimited	Five (5)
SUBSCRIPTION TERM	12 months	12 months
PHONE SUPPORT	Yes	Yes
EMAIL SUPPORT	Yes	Yes
WEB SUPPORT	Yes	Yes
REMOTE TROUBLESHOOTING VIA WEBEX/LYNC	Yes	Yes
PRIORITY 1 RESPONSE TIME SLA	1 hour	N/A
STANDARD RESPONSE TIME SLA	4 hours	4 hours
SUPPORT COVERAGE	All environments	Proof-of-Concept, Test & Dev.
RECOMMENDED FOR	Production Environments	Non-Production Environments
KNOWLEDGEBASE	Yes	Yes
ACCESS TO CUSTOMER SUPPORT PORTAL	Yes	Yes
UPGRADES AND PATCHES	Yes	Yes
EMERGENCY BUG FIX	Yes	No

To find out more about the Intel Distribution for Apache Hadoop software and partner program, visit **hadoop.intel.com**

Contact Intel at:
Phone: 1-855-229-5580
E-mail: asipcustomercare@intel.com