Extract Business Value from Big Data

Achieve Fast, Scalable ETL with Apache Hadoop* Running on Intel® Architecture

Organizations across public and private sectors are gathering and analyzing “big data” to forecast market conditions more accurately and to make better decisions about the issues that are most critical to their success. They sort through vast amounts of data from weather and economic reports, discussion forums, news sites, social networks, wikis, tweets, and blogs. They then analyze relevant data more deeply to gain new insights into their customers, operations, and competitors. Some even apply predictive analytics to identify the opportunities and risks they are most likely to face a month from now, a year from now, or even five years from now.

Yet big data introduces challenges as well as opportunities. Traditional Business Intelligence (BI) infrastructure cannot handle today’s massive, diverse, and fast-moving data streams. Apache Hadoop* running on Intel® Architecture provides a solution—a cost-effective, massively-scalable infrastructure that can ingest and store big data and prepare it for analysis. It provides a foundation you can implement today for targeted value, and then expand almost without limit to meet growing needs.
Break Through the Limitations of Traditional ETL

Today’s BI systems transform raw data into useful business information using a variety of sophisticated techniques, such as Online Analytical Processing (OLAP), data mining, process mining, complex event processing, business performance management, predictive analytics and prescriptive analytics. However, before you can analyze big data, you must first extract it from external sources, transform it to fit operational needs, and load it into an appropriate analytics environment—a process traditionally known as Extract, Transform, and Load (ETL).

Big data tends to overwhelm traditional ETL infrastructure. Incoming data streams are simply too large and fast-moving to be processed within acceptable time frames. The variety of data is also a challenge. Big data comes from many different sources, such as text documents, images, audio, video, operational logs, and sensors. These unstructured data types don’t fit well in a traditional relational database.

Apache Hadoop offers a solution to the ETL challenge. Based on technology developed by Google to power its widely-popular search engine, this open source software runs on scalable clusters of industry-standard servers configured with commodity storage devices. Through distributed storage and massively parallel processing, an Apache Hadoop cluster can scale to accommodate petabytes of poly-structured data.

Right-Size Your ETL Infrastructure for Greater Efficiency

ETL workloads vary, so a well-designed Apache Hadoop cluster is important to achieve your performance goals in the most cost-effective manner. Intel architecture offers a variety of options to help you implement a best-fit solution.

• **Cost-effective high performance for mainstream ETL workloads.** From a cost-benefit perspective, two-socket servers based on the Intel® Xeon® processor E5 family are the optimal choice for most Apache Hadoop workloads. These servers offer exceptional performance and are generally more efficient for distributed computing environments than large-scale multi-processor platforms. They also provide greater efficiencies in load-balancing and parallel throughput compared with smaller, single-socket servers.

• **Better cost models for lightweight ETL workloads.** Some ETL workloads, such as simple data sorting, cannot fully utilize the processing power of Intel Xeon processors. You can often run such lightweight workloads more efficiently on microservers based on the latest Intel® Atom™ processor. These server-class processors consume as little as 6 watts and offer valuable new data center efficiencies for applications that have sufficiently light processing needs.

Both Intel Xeon processors and Intel Atom processors support ECC memory, which automatically detects and corrects memory errors. Memory errors are one of the most common sources of data corruption and server downtime in the data center and a well-designed Apache Hadoop cluster has a lot of memory (typically about 64 GB or more per server). This large memory footprint increases the risk of errors, making ECC memory an essential feature.

---

Offload ETL with Hadoop

Using Apache Hadoop, organizations can ingest, process, and export massive amounts of diverse data at scale.
Storage performance is just as important as processing power in an Apache Hadoop cluster. Standard mechanical drives, in sufficient numbers, can meet many workload requirements. Intel® Solid-State Drives (Intel® SSDs) provide higher throughput with reduced latency. Intel tests show that replacing mechanical drives with Intel SSDs can increase cluster performance by as much as 80 percent.1

Network performance is also critical to ensure large data sets can be imported, processed, and exported at sufficient speeds. Intel 10 Gigabit Ethernet (10 GbE) server adapters offer a cost-effective, high-bandwidth option that scales easily to support cluster growth. As your cluster continues to expand, you can interconnect multiple 10 GbE switches and uplink to a faster networking infrastructure.

Reduce Your Operating Costs
In CIO surveys conducted by Gartner in 2007, 2010, and 2013, more than 70 percent of CIOs cite growing power and cooling requirements as the biggest challenges they face in their data centers.2 The exceptional energy-efficiency of Intel Xeon processors, Intel Atom processors, and Intel SSDs can help to reduce the load on your data center and your budget. Intel also offers an advanced power and thermal management application called Intel® Data Center Manager (Intel® DCM). Intel DCM takes advantage of instrumentation built into Intel® processors. You can use it to monitor and control power at all levels, from individual servers to entire facilities, so you can minimize energy consumption without compromising performance.

Reduce Your Risk
Open source Apache Hadoop software is available free of charge from the Apache Software Foundation. Enhanced software distributions are also available for a fee from value-added distributors, such as Intel. These enhanced distributions come with additional features and with service and support packages that help to simplify implementation and reduce risk.

The Intel Distribution for Apache Hadoop software (Intel® Distribution) is an open source product that includes Apache Hadoop and other components, along with enhancements and fixes from Intel (see www.intel.com/hadoop). This software is highly optimized for the latest Intel Xeon processors, Intel SSD storage devices, and Intel® 10 GbE network adapters. Tests show the combined platform delivers up to 30x higher performance than generic Apache Hadoop software running on a less optimized hardware platform.1

The Intel Distribution provides integrated support for key enterprise requirements, including:

- **Data confidentiality.** Hardware-accelerated encryption and granular access controls allow you to integrate sensitive data types without sacrificing security, compliance, or performance.

- **Scalability and availability.** Multi-site scalability and adaptive data replication simplify integration and help to ensure that critical data and insights are always accessible.

- **Advanced analytics.** Intel® Graph Builder and integrated support for R (an open-source application for statistical analytics) help data analysts and developers extract higher value from big data.

- **Service, support, and training.** Intel offers extensive online training and provides expert support for planning, implementing, and maintaining Apache Hadoop deployments based on the Intel Distribution.
Conclusion

Big data is generating new opportunities and new challenges for businesses across every industry. The challenge of data integration—incorporating data from social media and other unstructured data into a traditional BI environment—is one of the most urgent issues facing CIOs and IT managers. Apache Hadoop provides a cost-effective and massively scalable platform for ingesting big data and preparing it for analysis. Using Hadoop to offload traditional ETL processes can reduce time to analysis by hours or even days.

Running your Hadoop cluster efficiently means selecting an optimal infrastructure of servers, storage, networking, and software. Intel provides software as well as hardware platform components to help you design and deploy an efficient, high-performing Hadoop cluster optimized for big data ETL. Take advantage of Intel reference architectures, training, professional services, and technical support to speed up your deployment and reduce risk.

For more information about Intel big data solutions and resources, visit:

> hadoop.intel.com
> intel.com/bigdata
> intel.com/microservers