Why You Should Read This Document

This paper describes Intel’s perspective on the analytics of big data generated by sensors and devices on the edge of networks. The paper includes a discussion of:

- The importance of data at the edge of networks where some of “biggest” big data is generated
- How big data is inherently different from the data managed by traditional data management or business intelligence platforms, and why it matters
- A quick overview of emerging technologies, including distributed frameworks such as the Apache Hadoop* framework and Apache* MapReduce
- Four analytics use cases for government, retail, automotive, and manufacturing—two utilizing the Hadoop* framework and two focused on intelligent systems
Vision Paper
Distributed Data Mining and Big Data
Intel's Perspective on Data at the Edge

Sponsors of Tomorrow: Intel
Contents

3 Data at the Edge: New Opportunities for Big Data

4 Big Data and Emerging Technologies: The Abridged Version

5 Big Data at the Edge: A Closer Look

7 Harnessing Data from Intelligent Systems and Sensors

9 Use Cases for Data at the Edge

12 What’s Next?
Data at the Edge: New Opportunities for Big Data

The explosion of big data is testing the capabilities of even the most advanced analytics tools. IT is challenged by the sheer volume, variety, and velocity of this flood of complex, structured, semistructured, and unstructured data—which also offers organizations exciting opportunities to gain richer, deeper, and more accurate insights into their business.

The tremendous opportunities to gain new and exciting value from big data are compelling for most organizations, but the challenge of managing and transforming it into insights requires a new approach to analytics that has a far reaching impact on IT infrastructure. Traditional systems are unable to cope cost-effectively—if at all—with new dynamic data sources and multiple contexts for big data. Emerging technologies such as the Hadoop® framework represent completely new approaches to capturing, managing, and analyzing big data. Big data challenges plus new technologies are causing a paradigm shift that is driving organizations to reexamine their IT infrastructure and analytics capabilities.

Intel Perspective: The Importance of Data at the Edge

Intel believes that realizing the promise of big data analytics requires capturing and processing data where it resides. This paper explores the value of data at the edge of networks, where some of “biggest” big data is generated. As the use of sensors and devices as well as intelligent systems continues to expand, the potential to gain insight from the flood of data from these sources becomes a new and compelling opportunity. Businesses that can harness the power of big data at the edge and unlock its value to the organization will outperform their competitors with greater capabilities to innovate creatively and solve complex problems whose solutions have been out of reach in the past.

What Is Big Data?

Big data is typically described by the first three characteristics below—sometimes referred to as the three Vs. However, organizations need a fourth—value—to make big data work.

- **Volume.** Huge data sets that are orders of magnitude larger than data managed in traditional storage and analytical solutions. Think petabytes instead of terabytes.
- **Variety.** Heterogeneous, complex, and variable data, which are generated in formats as different as e-mail, social media, video, images, blogs, and sensor data—as well as “shadow data” such as access journals and Web search histories.
- **Velocity.** Data is generated as a constant stream with real-time queries for meaningful information to be served up on demand rather than batched.
- **Value.** Meaningful insights that deliver predictive analytics for future trends and patterns from deep, complex analysis based on machine learning, statistical modeling, and graph algorithms. These analytics go beyond the results of traditional business intelligence querying and reporting.
Big Data and Emerging Technologies: The Abridged Version

Big data management is inherently different from traditional relational models of data management or business intelligence platforms. While that difference is often described in terms of the data, “structured versus unstructured,” this isn’t quite accurate. Log data, for example (a growing source of big data), has structure. The difference is better described this way: Unlike relation-based data, big data manages data in any format and does not require the time and effort to create a model first to capture, process, and analyze your data.

<table>
<thead>
<tr>
<th>Application</th>
<th>Relation-Based Data</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data processing</td>
<td>Single-computer platform that scales with better CPUs; centralized processing</td>
<td>Cluster platforms that scale to thousands of nodes; distributed processing</td>
</tr>
<tr>
<td>Data management</td>
<td>Relational databases (SQL); centralized storage</td>
<td>Nonrelational databases that manage varied data types and formats (NoSQL and HBase* databases); distributed storage</td>
</tr>
<tr>
<td>Analytics</td>
<td>Batched; descriptive; centralized</td>
<td>Real-time; predictive and prescriptive; distributed analytics</td>
</tr>
</tbody>
</table>

Distributed Frameworks: The Apache Hadoop* Framework and MapReduce

New technologies are emerging to make big data analytics possible and cost-effective. The Apache Hadoop* framework is evolving as the best new approach. The Hadoop framework redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources.

The Hadoop open-source framework uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management. In addition to offering high availability, the Hadoop framework is more cost-effective for handling large, complex, or unstructured data sets than conventional approaches, and it offers massive scalability and speed.

MapReduce, the software programming framework in the Hadoop stack, simplifies processing on large data sets and gives programmers a common method for defining and orchestrating complex processing tasks across clusters of computers. MapReduce applications coordinate the processing of tasks for a cluster node by scheduling jobs, monitoring activity, and reexecuting failed tasks. Input and output are stored in the Hadoop Distributed File System (HDFS*). Typically the data is processed and stored on the same node, making it more efficient to schedule tasks where data already resides and resulting in high aggregate bandwidth across the node. For a more detailed look at the Hadoop framework and MapReduce, visit intel.com/bigdata.
Big Data at the Edge: A Closer Look

Much of the current discussion about big data analytics today focuses on managing and analyzing unstructured data from business and social sources such as e-mail, videos, tweets, Facebook* posts, reviews, and Web behavior. While this type of big data analytics promises to provide significant value to organizations, data generated at the edge of the network from sensors and other devices represents another huge, untapped resource with the potential to deliver insights that can transform the operations and strategic initiatives of public and private sector organizations.

Data from intelligent systems and sensors is some of the largest volume, fastest streaming, and/or most complex big data. The data sources are distributed across the network and data is collected by an enormous variety of equipment, such as utility meters, traffic and security cameras, RFID readers, factory-line sensors, fitness machines, and medical devices. Ubiquitous connectivity and the growth of sensors and intelligent systems have opened up a whole new storehouse of valuable information. Edge data can provide significant value to both the private and public sector as a source of enormous potential for gaining deeper, richer insight faster and more cost-effectively than in the past. In many cases, analysis of edge data can help organizations respond to events and solve problems that were previously out of reach.

Sensors from multiple systems at the edge stream data via the Internet, LANs, WANs, and mobile networks.
For an example of the size and scope of edge data, consider the machine-generated data from the engines of a Boeing® jet. Each engine generates 20 terabytes (TB) of sensor data every hour, so that a four-engine jumbo jet quickly reaches 640 TB of data during an Atlantic crossing. With more than 25,000 commercial flights in the U.S. sky on any given day, a single day of sensor data can measure in exabytes.¹

Humans are also generating sensed data. Deb Roy, director of the Cognitive Machines Group at the MIT Media lab, tracked the activities and sounds in his home for three years, starting with the day he brought home his newborn son. Analysis of more than 90,000 hours of video and 140,000 hours of audio mapped his son’s acquisition of speech and has provided enormous insight into how humans develop and learn.²

Harnessing Data from Intelligent Systems and Sensors

Clearly, the scope of big data on the edge is enormous. With the number of connected devices projected to reach almost 15 billion by 2015, the volume, variety, and speed of data generated from intelligent systems and sensors will be ever increasing. How can organizations harness and make sense of this fast moving data stream?

Leveraging Sensed Data and Grid Infrastructure

Big data at the edge is generated by embedded sensors and actuators in physical objects, which are linked through wired and wireless networks—often using the same communications protocol that connects to the Internet. This process of generating and analyzing data from intelligent systems and sensors is often referred to as the Internet of Things (IoT).

IoT is a major source of sensed data. Huge volumes of sensed data flow over the network to local computers or the cloud for analysis, and generate the intelligence for actuators to exert control over the physical world. Using MapReduce, the data is captured and processed where it resides at the edge by the local node and then sent wherever it is needed. In the case of an actuator, the results provide instant feedback that enables the device to modify activity.

Internet of Things (IoT)
At a Glance

- Instrumentation for sensing
- Intelligence for processing and control
- Interconnection for communication

3 “Global Internet Traffic Projected to Quadruple by 2015” The Network (press release) (June 1, 2011).
Implications for Technology

For data to be analyzed where it resides, compute and storage capabilities must be local at the edge and in the cloud. This local infrastructure must address a set of unique challenges based on characteristics of the data and related issues.

- Sensed data is massive and streams 24-7.
- Data is noisy and dirty and requires preprocessing.
- Data has strong locality characteristics, meaning that the devices are operated and consumed locally.
- Data ownership, interoperability, security, and privacy are big issues.

How does this translate into a real-life example? Here’s a transportation and public safety example.

- Road sensors may belong to different departments.
- Some cameras are owned by public security, while others belong to public transportation.
- Data is generated on private vehicles.

The issues: Can the data from these multiple systems be integrated and analyzed for meaningful insight? Who owns the data generated on private vehicles? Is the data secure?

These issues are well worth resolving. Multiple data streams can unlock intrinsic correlations that can have great significance overall. A recent study in a city in the People’s Republic of China (PRC) shows that if you can detect morning wash time from the water supply subsystem, you can infer the morning rush hour; similarly, if you can detect when offices are powered down in the evening, you can infer the evening rush hour. Understanding these relationships can help cities better handle traffic at peak times as well as improve availability of water and electrical resources when they are most needed.

For the hundreds of petabytes of data generated by intelligent systems and sensors, it’s too expensive and inefficient to move them to a central cloud. Plus, a central cloud is challenged to deliver real-time intelligence for the edge. Back to our road sensor example: The front end can’t wait for the central cloud to decide whether a car is running red lights.

Immediacy: How Fast Do You Need Insight?

Does every insight need to be in real time for organizations to drive value from their data? Actually, not all usage scenarios require real-time analytics. While applications at the edge may require immediate feedback to adjust equipment, insights based on the aggregation of that data may not be needed as quickly. Near real-time, near-line (periodic batching), or even batch processing may be timely enough.

Today, organizations in emerging markets are more likely to implement the Hadoop® framework to process both relation-based and unstructured data. More mature markets, such as Europe and the United States, already have traditional data management systems in place and are more likely to begin their forays into big data analytics with batch and near-line analytics. Ultimately, companies—even big Internet companies—will evolve to use a combination of real-time, near real-time, near-line, and batch processing for big data.

Intelligent Connected Systems

IDC described intelligent systems as those enabled with high-performance microprocessors, connectivity, and high-level operating systems. Embedded processors no longer perform as fixed functions that stand alone, but pack compute performance and integration into devices that fuel intelligent systems. Combined with cloud-based applications and analytics capabilities, these intelligent systems can derive value from edge data and bring the Internet of Things (IoT) to reality.

Use Cases for Data at the Edge

Technology already exists to enable organizations to build out architectures that can support distributed frameworks for big data and the local requirements of edge data. High-performance processors, 10 gigabit Ethernet solutions, and inexpensive storage options can support the clusters that run the Hadoop stack. The evolution of disparate embedded systems to intelligent connected systems continues to gain momentum, as does the development of cloud and big data analytics platforms. Understanding use cases for data from intelligent systems and sensors will further the development of customer requirements, standard architectures, and end-to-end, interoperable solutions that enable architectures for those scenarios. The value to organizations extends across most industries. The following are examples of use cases from four of them: government, retail, automotive, and manufacturing.

Smart Cities: Improved Urban Performance

“Smart city” is a concept that describes the use of a smart-grid infrastructure (physical capital and information and communications technologies) to improve environmental sustainability, manage energy consumption, better coordinate public resources, protect the quality of life for urban and metropolitan citizens, and plan for sustainable growth.

Intel is currently involved in smart city innovation projects in the United States, Europe, and the PRC. These initiatives are exploring how intelligent systems at the edge can improve the management of the urban environment. For example, utility companies and governments are using data from the smart grid to understand the complex relationships between generation, distribution, and consumption—with the goal of delivering reliable energy and reducing operating costs. At the same time, consumers are able to use data from the smart grid to better manage their personal energy requirements; for example, a “not-home” state might turn off lights, shut down unused equipment, and adjust temperature.

By implementing intelligence throughout the electrical network, grid devices and compute nodes are enabled with capabilities to measure, analyze, and predict. Optimized decisions are made closer to the edge rather than only at a centralized control center. Communication between devices helps determine when, where, and how much energy should be produced, and consumers can use home management tools to monitor and adjust energy consumption.

Intel Science and Technology Center (ISTC) for Big Data

Intel is supporting big data research at the newest Intel Science and Technology Center (ISTC) at MIT’s Computer Science and Artificial Intelligence Laboratory. The research will seek ways to accelerate the pace of big data innovation in diverse fields such as government, financial services, healthcare and life sciences, manufacturing, and retail. With MIT as the hub, collaborators will include faculty from the University of California at Santa Barbara, Portland State University, Brown University, University of Washington, and Stanford University.
Retail: Connected Stores

What if retailers could know their customers like Amazon.com knows its customers? Traditional merchandizing systems collect point-of-sale data and aggregate it by store, district, region, time, and product categories but lose the insight that detailed transactions can give when tied to a specific customer. Plus, the size of the data, the level of detail, or the cost may prevent retailers from storing SKU details for more than a few months. The Hadoop framework changes the economics of this by radically lowering the cost of storing data and more accurately market to individuals rather than a demographic.

Retailers are using a variety of intelligent connected systems that gather data and provide immediate feedback to help them to engage shoppers, including:

- **Digital signage** to measure advertising effectiveness, adapt messages to specific audiences, and provide highly individualized information
- **Transaction and point-of-sale systems** that provide product availability, suggest complementary purchases, and drive up-sell
- **Intelligent vending machines** that engage passersby with interactive displays, video analytics, digital signage, and cashless display systems to dispense everything from samples for feedback on new product ideas to fresh food and upscale accessories such as jewelry

- **Interactive kiosks** that bridge to online or in-store environments and use shopper profiles or past purchase data to offer suggestions or directions to items of interest, either online or in the physical store
- **Digital security surveillance** that prevents theft, locates lost children, and gathers demographics such as the traffic in key store areas to assist merchandising efforts

The ability to gain insights from the data generated by these systems makes it possible to provide customer-centric “connected stores” and provide an “on-store” environment that weaves together online and in-store operations for an emotionally satisfying experience for shoppers, while at the same time optimizing operations. Customers can find what they want faster from trusted retailers across multiple channels and engage at any stage of the buying process. Retailers can integrate their supply chain activities with actual shopper behavior and deliver a great shopping experience consistent across all touch points with a specific customer. Plus, they can provide customers with opportunities to engage with their brands in more meaningful ways to cement customer loyalty.
Automotive: Connected Intelligence on the Road

The convergence of IT and the consumer experience in automotive is growing rapidly in the form of intelligent in-vehicle systems. These systems are transforming the in-car experience by enabling the seamless connection between vehicles and connected devices, including consumer electronics, mobile devices, and sensors. In addition to streaming video for the kids, these data sources can be aggregated and analyzed to provide immediate insights—for example, location data could be combined with road work and other traffic information to help commuters avoid congestion or take a faster route. Other applications of big data could be used to help:

- Monitor driver alertness and look for signs of medical distress with built-in cameras that use facial recognition software. The system can read expressions and automatically sound audible alarms, stop the car safely, and contact emergency services as necessary.
- Connect drivers and passengers with friends by providing notifications when they may be nearby.
- Provide alerts about upcoming signage, blind spot obstacles, and road conditions.
- Proactively monitor vehicle operating conditions and warn of potential malfunctions in advance.
- Offer valuable new automotive services and applications that improve both customer relationship management and vehicle relationship management.
- Use a smartphone for remote keyless entry or to provide alerts about tampering or impact.
- Detect real-time traffic flow from each direction and automatically change traffic signals to improve flow.
- Enable automated, intelligent, real-time decisions to optimize travel across the transportation infrastructure, as cars become capable of connecting to the roadway, safety systems, and one another.

Manufacturing: Smart Factories

Information technology and operations technology are converging in unprecedented ways in smart factories. While most factories today are highly automated, they are purpose-built for specific production processes. Device and control layers on the factory floor can’t exchange information with the business and data networks that run the company.

Smart factories, on the other hand, connect the boardroom, the factory floor, and the supply chain for higher levels of manufacturing control and efficiency. Sensors and actuators in devices such as cameras, robotic machines, and motion control equipment generate and use data to provide real-time diagnosis and predictive maintenance, increased process visibility, and improved factory uptime and flexibility.

Specific usage scenarios include:

- Communication across the factory floor and with enterprise IT systems for more efficient coordination of plant resources, employees, and suppliers
- Detection of failure conditions to enable faster response
- Greater situational awareness, seamless multizone protection on the factory floor, native supervisory control and data acquisition (SCADA) support, and remote device management
- Robotics that dramatically improve productivity and industrial safety
- Monitoring production line activities for product quality issues
What’s Next?

Big data is a game changer—and it’s already here. While most of the momentum around big data today is around social media sources, Intel believes that realizing the promise of big data analytics must include a way to harness the potential of big data from intelligent systems and sensors.

Intel sees the following next steps as critical for organizations who want to take advantage of edge data sources:

- Understand use cases and their implications. We must understand how existing disparate data sources can be evolved into a network of integrated, intelligent, connected systems.
- Define the usage model requirements for the analytics of edge data. The architecture must take advantage of big data distributed frameworks to move computation closer to where the data resides and support big data analytics at the edge via intelligent systems and local clouds.
- Enable the fast and secure delivery of aggregated data from edge analytics systems to other cloud and analytics platforms for further analysis.
- Address issues related to data ownership, interoperability, security, and privacy.

As interest in data from intelligent systems and sensors continues to grow and organizations understand better how they can use it, Intel is at the forefront of this emerging topic. Intel is already taking a leadership role with cloud computing and big data analytics.

As technical advisor to the Open Data Center Alliance (ODCA), an independent IT consortium comprising global IT leaders from more than 300 companies, Intel will play a major role in the newly formed Data Services Workgroup as it works to define usage model requirements that support the secure collection, management, and analysis of big data; drive benchmarking for the Hadoop framework; and develop interoperable standards that make big data frameworks cloud ready.

Plus, Intel has years of experience providing the technology that powers intelligent systems, as well as platforms that deliver the exceptional performance, low latency, and high throughput needed to handle large data sets and transform them into deep insights.

Count on Intel for the technology, guidance, and vision to make big data work for you.

Take the Next Steps to Manage and Analyze Edge Data

Here's how you can get ready to take advantage of this fast moving area for your organization.

- Keep up-to-date with what's happening. Intel offers practical guidance to help you deploy big data environments more quickly and with lower risk. Go to intel.com/bigdata.
- Explore business opportunities deriving from the analytics of edge data. Collaborate with the business to understand existing edge systems and the potential use for data. For more information, go to intel.com/intelligentsystems.

To learn more about edge data and big data analytics, visit the IT Center at intel.com/bigdata.