

IT@INTEL

Using Big Data to Understand the Impact of Email on Business

Using the results of our email study, we plan to work with business units to help employees manage email better and reduce the impact of email on business.

Tridib Chowdhury
Enterprise Architect, Intel IT

Kiran Ramineni
Big Data Application Architect and
Technical Lead, Intel IT

Nayem Rahman
Big Data Application Developer
Lead, Intel IT

Meghana Krishnan
Lead BI Developer, Intel IT

Sunita Sharma
Big Data Capability Engineer, Intel IT

June Wong
Big Data Project Manager, Intel IT

Executive Overview

Intel IT is responsible for fueling Intel's business transformation with information technology, which includes providing employees with the ability to effectively communicate and collaborate. While email is pervasive in most businesses today, it can siphon time, productivity, and creativity from the workforce. Intel IT decided to explore the correlation between productivity and email. We conducted a study to determine email usage, while complying with Intel's policies regarding employee privacy rights.

To conduct the study of the impact of email on business, we created a custom architecture using Apache Hadoop*, Hive*, Impala*, Pig*, and Sqoop*. Based on four months of data collection, we were able to identify the highest email users at Intel by region, Intel site, and business unit. We were also able to determine a baseline email volume, recognize trends, and identify usage patterns and the main sources of email.

We plan to coach groups with the heaviest email use about proven methods for managing and reducing the time spent on email, thus enabling them to focus on higher-value activities. We will also work with the business units to determine whether the use of collaboration tools or corporate web portals might be more effective for teamwork and communication.

Contents

- 1 **Executive Overview**
- 2 **Background**
- 2 **Pilot Study**
 - Maintaining Employee Privacy
 - Solution Architecture
 - Extract, Transform, and Load
 - Front-End Reporting
- 8 **Results**
- 9 **Next Steps**
- 10 **Conclusion**

Contributors

Mark Densham

Director of IT Collaboration and Productivity Engineering, Intel IT

Mandanna Manotira

IT Service Owner of Messaging Services, Intel IT

Xiongxin Lin

Business Intelligence Analyst, Intel IT

Lleyana Jack

Business Intelligence Analyst, Intel IT

Acronyms

BI	business intelligence
CDH	Cloudera Distribution of Apache Hadoop*
CDIS	Computer Design and Integration Systems
ETL	extract, load, and transform
HDFS	Hadoop Distributed File System*
HiveQL	Hive* query language
ODBC	open database connectivity

Background

Many industry studies have examined the productivity impacts of email. One study showed that employees at medium and large companies receive an average of 304 business emails weekly. These employees tend to check their email 36 times an hour. It takes the same employees about 16 minutes to refocus on work tasks after handling the incoming email. This study estimated that email's annual productivity costs per employee was USD 1,250 for reading spam and USD 1,800 for reading unnecessary emails.¹

The Radicati Group 2014 report showed that business email accounts for over 108.7 billion emails sent and received per day.² Contatta research found North Americans spend 75 billion hours each year on email, costing businesses USD 1.7 trillion in lost productivity.³

In addition to consuming a significant amount of time, email distracts employee from other job activities. Some surveys have discovered that people check for new emails 41 times a day.⁴ This equates to employees having their attention diverted 41 times during the workday instead of being able to focus on the tasks at hand.

While these studies show the cost of lost productivity, there is also a cost to the data center. Email consumes network bandwidth and requires storage resources and email servers. However, our main concern is helping Intel employees be more productive and understanding how employees consume and disseminate information so we can provide them with the tools they need.

Pilot Study

For a corporation that has used email communication as the primary communication vehicle for over 20 years, reducing email volume is a challenging goal. Intel IT conducted an email study to understand how employees consume and disseminate information, with the goal of reducing email volume corporate-wide. To achieve our objective, we set out to do the following:

- Obtain a baseline email volume.
- Recognize trends.
- Identify patterns of usage and the main sources of email.

Having this information would enable more efficient governance and allow us to propose better work methods or solutions to increase productivity using alternatives to email.

¹ Clark, Thor. "The Real Cost of Productivity" (Infographic) Blog. September 24, 2014. blog.barracuda.com/2014/09/24/the-real-cost-of-productivity-infographic

² The Radicati Group, Inc. "Email Statistics Report, 2014–2018." April 2014. www.radicati.com/wp/wp-content/uploads/2014/01/Email-Statistics-Report-2014-2018-Executive-Summary.pdf

³ Contatta, Inc. "The Staggering Cost of Business Email" (Infographic). May, 2014. blog.contatta.com/the-staggering-cost-of-business-email-infographic/345

⁴ Ron Friedman, PhD. "The Best Place to Work: The Art and Science of Creating an Extraordinary Workplace." 2014. Perigee Books.

Maintaining Employee Privacy

In many business organizations, employee access to email and the Internet on the corporate network can create potential liability issues and raises concerns about employee productivity and the downloading of inappropriate information.

An employee's right to privacy in the workplace, including the topic of workplace monitoring, is an increasingly sensitive legal topic, especially in an age of increased reliance on computers and electronic mail to do business. Laws and regulations around the world are often inconsistent and rapidly evolving, particularly in Europe. While companies have an obligation to protect their network assets and intellectual property assets as well as their employees' privacy, many jurisdictions also provide a privacy right to employees, including email communications details such as sender, receiver, and content. Companies must be mindful of privacy laws and regulations and how they may impact employee communications.

We developed and implemented a Privacy Plan for this project to verify that the actions we took complied with policy and applicable laws and regulations.⁵

Employee emails may contain personal health, financial, or family information that is viewed as sensitive in most jurisdictions, adding a layer of complexity to privacy compliance. In addition to protecting personal information, we needed to protect and provide security for internal business-confidential information. Because of these concerns, we needed to carefully collect and handle the data according to our approved Privacy Plan. In accordance with Intel's corporate privacy program, this study respected employee's privacy, did not query any individual's mailbox or content, and masked individual user data. Further we established strict security guidelines, designating which IT employees had access and authorization to work with the study's metadata and Apache Hadoop* architecture.

Solution Architecture

Intel uses email extensively between business units, suppliers, and customers. In our study, the email data amounted to approximately a billion database rows after just four months of downloading email server logs. The size of this data categorized it as a high-data-volume environment—this data grows by tens of millions of rows every month. We knew we needed a scalable solution to establish a baseline email volume. Instead of depending on traditional data extraction and storage capabilities, our approach was to build an architecture to host and store the data on a highly scalable platform using the Cloudera Distribution for Apache Hadoop (CDH) and a Hadoop Distributed File System* (HDFS*).

⁵ For more information about Intel's stance on privacy, see the white paper "Applying Privacy Principles in a Rapidly Changing World."



How Intel Developed Its Privacy Program

"As new technology and data uses challenge traditional privacy guidance, Intel recognizes the enduring value of this guidance and seeks ways to implement recognized privacy principles flexibly and effectively.

"Dramatic, fast-paced developments in digital technology present unprecedented challenges for companies applying privacy law, regulation, and commonly accepted principles. As a result, traditional notions of fair information practice principles that form the basis of privacy guidance are under scrutiny. Intel believes businesses and policymakers should not discard these enduring principles, but instead consider new implementations that effectively protect privacy and encourage innovation."

Read more in the Intel white paper [Applying Privacy Principles in a Rapidly Changing World](#).



In our study, email data amounted to approximately a billion database rows after just four months of downloading email server logs.

Three Solution Objectives

- Obtain a baseline email volume.
- Recognize trends.
- Identify patterns of usage and the main sources of email.

Using only Hadoop* ecosystem tools and technologies for ETL purposes, as well as for provisioning data to reporting tools, is a significant paradigm shift.

Similarly, we knew we would be unable to achieve insight and visibility into the data using standard operational reports and dashboards, or traditional data warehousing techniques. The performance of the reporting solution was also a primary concern, because of the volume of data being generated. We also needed a reporting solution that could display data in a manner that facilitates decision making, while maintaining employee privacy and data accuracy.

We wanted to develop a visualization component that could help us recognize trends and identify the main email sources and usage patterns. By using high-performance middleware components compatible with Intel's business intelligence (BI) platform, we could avoid data hops and access the data directly in Hive* tables. We tested the Impala* ODBC (open database connectivity) driver with Hive tables in the Hadoop data warehouse. Impala is Hadoop's high-performance engine, which enables faster querying of data and reduces roundtrips to the server. Another challenge we faced was the unstructured source email data, which is difficult to model in a schema format in a relational model for querying purposes.

BI derived from big data involves acquiring the data, then cleansing, formatting, and loading it into a data warehouse for analysis, a process known as ETL (extract, transform, and load). Using only Hadoop ecosystem tools and technologies for ETL purposes as well as for provisioning data to reporting tools is a significant paradigm shift, because it avoids storing data in conventional relational databases. Both Hive, which we used for ETL purposes, and Impala, which we used for reporting, used the same data sources that reside in the HDFS.

Hive is a MapReduce-based data warehouse. Hive provides a query language that is equivalent to a conventional SQL database. Hive Query Language (HiveQL) is run against data stored in the HDFS. HiveQL is converted to the Java* language to run as a MapReduce job. Both academic and industry papers have reported that MapReduce jobs are not efficient and take relatively longer to run than Impala queries do. However, we found Hive to be a good choice when it comes to dealing with a large volume of data in terms of extract, transform, aggregate, and load. We found that Hive also provides more fault tolerance compared to Impala. In the ETL process, fault tolerance helps maintain data integrity. Taking all these factors into consideration, we decided to use Hive for ETL purposes. We observed that the runtime of Hive ETL queries was reasonable. The queries took from one to four minutes to load 13 report tables.

Extract, Transform, and Load

We used Hadoop and associated tools to build a scalable ETL platform for our email study. Figure 1 shows the ETL process we used.

Extract Phase

For our analysis, we first captured the daily logs from the email servers, using Samba*. We wrote a custom script to initiate a connection between a Samba client and the email servers and extract the raw data. Computer Design and Integration Systems (CDIS) data was pulled daily using a Sqoop* connection to an SQL database. This pull model was managed in Linux*, which eliminated the need for virtual machines or a temporary staging area. Once the data was retrieved, it was split into two Hive tables. The Samba client generated log files during its attempt to connect to each email server. If the Samba client failed to connect to an email server, a failed server log file was generated (see [Data Quality and Integrity](#) for more details).

Intel's email servers are located worldwide. We excluded some email server daily logs from certain countries due to those countries' privacy laws. The extracted data was unstructured metadata. We saved only certain fields, such as the address of the receiving server, which server sent the email, message ID, event ID, and time stamp for the email. We knew some servers might be offline when accessing the data. Fail-safe logic recorded which servers were unavailable and should be retrieved again at a later time. The new retry data was appended in the Hive database.

Transform Phase

The data from the email servers was merged daily into a single table. We used 111 staging folders that were ultimately merged into one folder.

We used Pig* scripts to parse the source data, which was in CSV (comma-separated values) format, remove unwanted columns, and structure the data. Next, the structured data was partitioned and transformed to a Hive database, using Hive LEFT OUTER JOIN queries on the CDIS exclude table and daily email data. The database was secured using the Cloudera Sentry* service, which managed security and modularized the access controls. Multiple days of data were collected into Hive partitions.

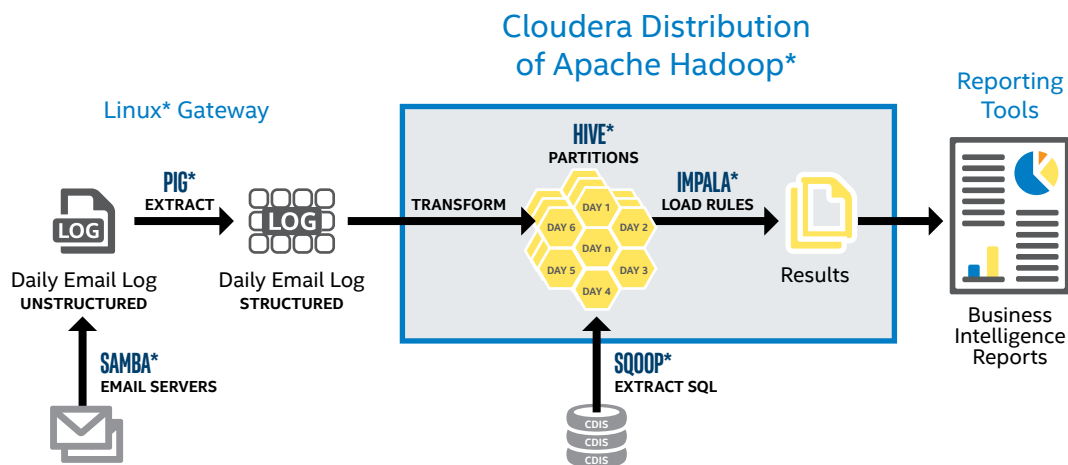


Figure 1. Hadoop* and associated tools enabled us to architect a high-performance extract, transform, and load (ETL) platform that supports data visualization and exploration.

In compliance with Intel's privacy program, we removed subject line information with the Pig script. Additionally, rules were applied for removing the sender name, IP address, recipient name, subject matter, and so on. The results were published to Intel's BI platform and made available through a self-service report.

A sample report from the analytics on the email volume had the following components:

- Number of emails sent per email server
- Number of emails sent by
 - Mailbox type (user, system, or generic)
 - Business group and organizational level
 - Geographic region and site
 - Message class (message, meetings, alerts, and notifications)
 - Employee type (Intel employee or contractor)
- Top 10 sources of email and mailbox classification
- Number of email recipients by distribution type

Load Phase

In this section, we explain how we moved the large amounts of Hive data into tables of structured data—that is, back-end reporting. The data was analyzed within a shared cluster. Through this architecture we eliminated the use of a relational database management system. We moved whole data sets to Hadoop and did the reporting from Hadoop, which enabled one less step of data staging. Using Impala enabled us to achieve rapid query responses and high performance. Also, generating reports directly through Hadoop reduced the number of required servers and thereby reduced the project's total cost of ownership.

Daily Reporting

The daily reporting batch job enabled us to segregate report data that could be ported and archived for trend analysis, problem persistence, and root cause analysis. This design reduces the time necessary to verify the data; verification was reduced from a few days to a few hours. By using aggregate queries on daily reports, we discovered a positive side effect: a boost in performance. For example, it took an average of only 95 seconds to run aggregated queries on four months of stored report data, using granular daily reports. Running the same query on the main Hive database with four months of all data partitions took about 500 seconds. That is, aggregated results using granular report tables not only returned exactly the same results as from direct Hive access, but produced a 5x improvement in performance.

ETL by the Numbers

Extracting and transforming Intel's email data was daunting in terms of data volume and processing.

- We captured logs from 111 email servers. The daily metadata log files from all these servers averaged 100 MB.
- Extracting the data from the servers took 2.5 hours—we did not have any parallel connections. It took another 2 hours to transform and load the data using Hadoop*.
- The raw source file was slightly more than 16 GB. We compressed the data using the Parquet indexing format and the Snappy compression codec.
- Data compression for Hadoop storage enabled us to save 90 percent of storage disk space—the compressed file size was only 1.6 GB.

While we were collecting the data we noticed that the access time was not proportional to the number of records, as shown in the following table. Even huge amounts of data could be accessed in a reasonable length of time.

Number of Rows	Access Time
48 million	37 seconds
568 million	39 seconds
1,635 million	48 seconds

Data Quality and Integrity

Even though both Hive and Impala used the same HDFS files as a data source, the ETL process in Hive did not cause data integrity issues during table loads, because data did not become available to Impala while tables were being loaded using Hive. Instead, data was made available to Impala once Hive ETL processes completed all table loads. If data had been made available to Impala after each table load, Impala might have found refreshed partial tables, causing reports to pull incomplete results. After all tables in a batch cycle were loaded, we took advantage of the HDFS metadata update command, issuing such a command for each individual table to make newly refreshed data available to Impala.

To check the data accuracy, we implemented an automatic email server log check. We pulled the message tracking logs from the email servers and added these logs to the Hadoop database. A flag was checked during the data pull if there was a successful connection to the email servers. If the connection was not successful, a failure log was generated. Once the logs were added to the database, we checked the connectivity to the servers listed in the failure logs. If the share permission in the analytic account for a server was not set properly, we notified the appropriate group. Automated unit tests, along with interactive spot tests, were added with Hive queries.

Front-End Reporting

Performance considerations and the use of visualization graphs made it challenging to architect an analytic solution on high-volume data. It was also challenging to maintain the validity and credibility of the data being calculated and presented. Further, using a star schema in a big data environment such as Hadoop can be quite different from using a traditional data warehouse.

Leveraging CDH's high-performance engine, the Impala ODBC driver serves as the middleware tool to connect to and query the back-end data stored in the HDFS. Using direct queries of the HDFS eliminates data hops from Hadoop to a staging database, and the BI platform can analyze data directly from the Hadoop data source. It further avoids using MapReduce algorithms which can significantly slow performance. Impala pulls data from the HDFS using a massively parallel processing architecture (many separate CPUs running in parallel) and executes the query operations in-memory.

Our solution architecture used semi-aggregated, de-normalized tables as a reporting layer schema within the Hive schema. We built a relational schema in a single front-end environment, based on date-time batch processing, to facilitate report building and querying. This

Opportunities for Technology Maturity

Based on our experience during the email study, we found that the Hadoop* open source tools such as Hive* have not quite reached an enterprise maturity level. We encountered a few limitations of Hive queries that presented challenges during development and testing, requiring us to create workarounds.

Hive uses an SQL-like language called Hive Query Language (HiveQL). One issue we encountered was that in some cases the Hive query executed with LEFT OUTER JOIN as INNER JOIN. This means that the query failed to pull the complete set of data from the left table. It filtered out all those records from the primary source table that were not available in the secondary source table. We were able to detect that data loss issue during our preproduction testing phase. Through trial and error, we found that adding the COALESCE function to join columns caused the query to correctly execute the LEFT OUTER JOIN. We suspect that HiveQL could not handle NULLs encountered in the join columns.

Another issue with HiveQL was that load queries sometimes loaded corrupted data in the table when we attempted to load data for more than one day (such as loading data for the month of February with date partitions). We noticed that HiveQL was not loading data for the first day of February. On the other hand, it was loading duplicate records for the last day of February. However, loading data for one day (one date partition) at a time worked. This issue compelled us to run HiveQL 180 times—one query for each date—to load data into the table worth four months.

Also, the Hive version we used (0.13) was not capable of executing delete and update operations. In our project sometimes we needed to delete a subset of data (such as for a particular date), but Hive was able to delete a subset of data only by partitions. This did not cause us undue difficulty because all of our tables were defined to store data by date partitions. Therefore, we were able to delete data by date partitions in this project. Hive version 0.14 has the capability to delete and update rows from Hive tables.

approach shifted the processing burden to a high-performance engine and enabled a positive user experience as well as data drill-down and manipulation without impacting data accuracy in the front-end layer.

Due to aggregations, most of the underlying report tables contained relatively small sets of data, except for one table that had more than 80 million rows. For this large table, the Impala queries took about one minute; the reporting layer was able to query data within a few seconds for the remaining tables. Our approach to front-end reporting enabled data visualizations on unstructured, high-volume data, with acceptable rendering times, in a front-end environment that was familiar to our users.

Results

Data was collected and sorted by business unit, site, and quarter. We gained a view into Intel's email traffic, identified specific business units and sites that are very active with email, and identified applications that generated the most email. The results also gave us insight into how active the email servers are on a particular day.

In general, the results of our analysis were not surprising. As shown in Figure 2, more than 45 percent of all email was generated by infrastructure applications (business applications and generic mailboxes). The next largest category was email to Intel employees from an external domain, followed by email between Intel employees.

Figure 3 shows email analysis by region; the majority of email originated in the United States. We drilled down in the U.S. data to determine which states generated the most email. As we expected, the states with the most Intel employees—California, Oregon, and Arizona—generated the most email.

Also as expected, the business units with the most employees generated the most email, as shown in Figure 4.

In addition to the factual findings, the study raised the following challenges:

- The interpretation of the data and business, legal, and privacy rules required many rounds of refinement.
- This project was the first time we had used Intel's BI platform with Hadoop. Setting up the architecture involved more technical and functional complexity than we expected.

This first study serves as our baseline, and we plan to continue to collect email statistics to monitor any future changes, which will help us further establish the validity of the data we have already collected.

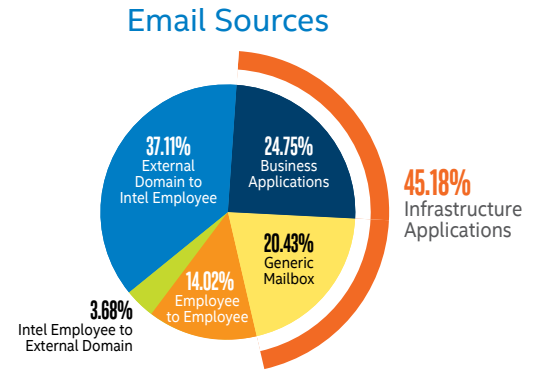


Figure 2. Our study gave us insight into the main sources of emails. More than 45 percent of all email was generated by infrastructure applications (business applications and generic mailboxes).

Total Messages Sent

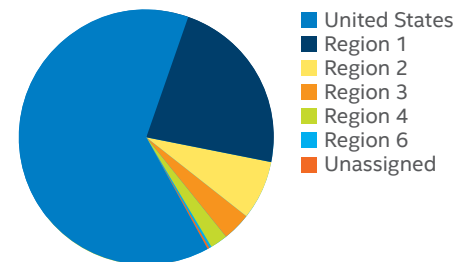


Figure 3. We determined that the majority of email originated in the United States.

Business Unit Messages (in Millions)

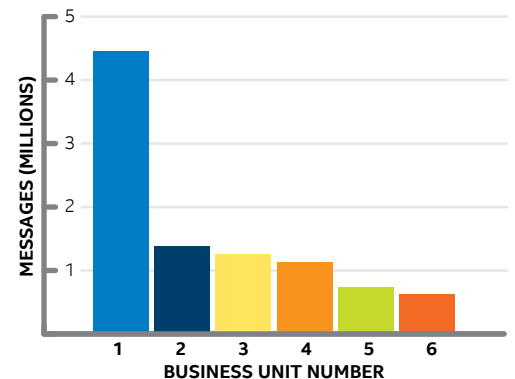


Figure 4. Business unit 1, which has the most employees, generated the most email.

Next Steps

We plan to share the results of this study with Intel's CIO. In this review, we will discuss the business impact of email and explore collaboration tools that could have less productivity impact.

We also plan to work with a few business units that generate the most email. For example, we identified one business unit that has few employees but generates a disproportionately large share of email. We want to understand how this group uses email and estimate its productivity impact. We hope to discuss with them how they might improve productivity and reduce email use by taking advantage of other collaboration tools, such as social sites, and workflow and content solutions, that can help them work more efficiently. Based on these discussions, we plan to model what is successful and implement that with additional business units with high email usage—resulting in employees being able to work on tasks with higher returns.

These conversations will also include email reduction techniques that business groups could use to improve their employees' productivity. For example, employees might consider implementing the following guidelines:

- Track how many times they check their email every day for one week. Being conscious of email habits can help them become more disciplined and focused.
- Limit emails to five sentences or less. If a longer response is required, call the person. This helps build work relationships and minimizes miscommunications.
- Refrain from checking for new email unless they have the time and energy to deal with the contents. Checking email when tired or in the middle of another project wastes time.
- Stop organizing email into folders. Searching through the inbox can be faster than setting up folders and sorting emails.
- Turn off email alerts. Email alerts interrupt thought processes.
- Turn off the email system. Even an hour-long “holiday” from email can help them focus on higher-value tasks.
- Unsubscribe with abandon. If it is not important enough to look at right now, delete it.
- Touch an email once and take action immediately, if a response is possible in two minutes or less. For example, reply immediately or immediately send a meeting request.

We also plan to export a portion of aggregated data to an SQL server for advanced analytics and data visualization. This process will provide an almost real-time dashboard to analysts, executives, and administrators. The next email analytics study will expand the data set, integrating information about meetings and usage of online meeting resources, conference rooms, and video. We hope this next study will enable us to establish a correlation between collaboration tools and diagnose employee behavior regarding solution adoption and benefits.

The next email analytics study will expand the data set, integrating information about meetings and usage of online meeting resources, conference rooms, and video.

Conclusion

In pursuit of our goal to enhance employee productivity, Intel IT conducted a study that forms the foundation for reducing email usage across the enterprise. We suspected Intel had a high volume of email, which according to external studies impedes productivity. Email can often be replaced by better communication solutions. Our study, while protecting employee privacy, revealed which Intel sites and business groups generate the most email.

Email analytics is a big data use case common to many organizations. Collecting email logs results in a constantly growing data set—requiring a scalable solution. We set up a custom architecture, built using the Hadoop ecosystem of tools, to capture the logs from Intel's email servers. We were able to analyze email without violating privacy or legal requirements. The data proved our hypothesis correct—Intel's email volume was significant. Further, we found email traffic generally aligned with the size of Intel facilities and business groups, with one notable exception.

We plan to work with business groups to help them manage email better and reduce the impact of email on business. We will use our study results to provide recommendations to help Intel improve efficiency and to guide future deployment and use of collaboration and messaging services.

For more information on Intel IT best practices, visit www.intel.com/IT.

Receive objective and personalized advice from unbiased professionals at advisors.intel.com. Fill out a simple form and one of our experienced experts will contact you within 5 business days.

IT@Intel

We connect IT professionals with their IT peers inside Intel. Our IT department solves some of today's most demanding and complex technology issues, and we want to share these lessons directly with our fellow IT professionals in an open peer-to-peer forum.

Our goal is simple: improve efficiency throughout the organization and enhance the business value of IT investments.

Follow us and join the conversation:

- [Twitter](#)
- [#IntelIT](#)
- [LinkedIn](#)
- [IT Center Community](#)

Visit us today at intel.com/IT or contact your local Intel representative if you would like to learn more.

Related Content

Visit intel.com/IT to find content on related topics:

- [Transforming Sales Productivity with Social Collaboration white paper](#)
- [Applying Privacy Principles in a Rapidly Changing World white paper](#)
- [Aligning Social Collaboration with Business Objectives white paper](#)
- [Evaluating Apache Hadoop* Software for Big Data ETL Functions white paper](#)
- [How Intel IT Successfully Migrated to Cloudera Apache Hadoop* white paper](#)



THE INFORMATION PROVIDED IN THIS PAPER IS INTENDED TO BE GENERAL IN NATURE AND IS NOT SPECIFIC GUIDANCE. RECOMMENDATIONS (INCLUDING POTENTIAL COST SAVINGS) ARE BASED UPON INTEL'S EXPERIENCE AND ARE ESTIMATES ONLY. INTEL DOES NOT GUARANTEE OR WARRANT OTHERS WILL OBTAIN SIMILAR RESULTS.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS AND SERVICES. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS AND SERVICES INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others. Copyright © 2015 Intel Corporation. All rights reserved.

Printed in USA

Please Recycle

1015/LMIN/KC/PDF