WHITE PAPER
OCTOBER 2015

IT@INTEL

How Intel IT's Integrated Analytics Platform Helps Sales and Marketing

David Schaefer
Business Intelligence Architect, Intel IT

Ivan Harrow
Director of Insights & Analytics, Intel IT

Greg Martinez
Program Manager, Intel IT

Gonzalo Lopez
Application Developer, Intel IT

Seshu Edala
Application Developer, Intel IT

Executive Overview

Intel IT is building an analytics platform that integrates and connects business intelligence (BI) data using a data lake model to reduce insight latency from months to days. Eventually, we aim to reduce data insight latency to 24 hours.

Using the Integrated Analytics Hub, data analytics projects have already accounted for an estimated quarterly savings on marketing digital-media expenditures of approximately USD 170,000.

A data lake is a repository for large quantities of structured and unstructured data. It allows for more versatility than a data warehouse because we can ingest data from various streams at various rates and according to various data models and file formats.

Intel's sales and marketing organizations rely on BI solutions from Intel IT to help make smarter data-driven business decisions. Analytics solutions in environments that are incapable of interconnecting datasets weaken the business value we can provide sales and marketing.

The Integrated Analytics Hub (IAH) delivers the following benefits to sales and marketing:

- By implementing a data lake model using Cloudera's distribution of Hadoop* (CDH), we provide data scientists, analysts, data stewards, and end users faster and more flexible access to large volumes of data. The data is available in multiple formats and in three states—raw, cleansed, and conformed—for various levels of analysis.
- All users can more easily engage in self-service BI and use BI front-end tools of choice for analysis of data at any of the three states.
- The ability to interconnect datasets and share visualizations, reports, and dashboards on the self-service BI portal increases velocity and removes manual IT intervention.

Using IAH, data analytics projects have already accounted for an estimated quarterly savings on marketing digital-media expenditures of approximately USD 170,000. We continue to optimize IAH to increase automation and develop applications that can simplify and accelerate self-service BI.

Contents

Contributor

Rich Mason, Systems Analyst, Intel IT

Acronyms

BI	business intelligence
CDH	Cloudera's distribution of Hadoop
CRM	customer relationship management
ETL	extract, transform, load
HDFS	Hadoop Distributed File System
IAH	Integrated Analytics Hub

Business Challenge

For Intel's sales and marketing organizations, integrated data analytics has the potential to provide transformative insights that can increase revenue opportunities. However, for sales and marketing to derive optimal business value from this insight, integrated analytics must match the velocity of sales and marketing's business needs. For results to be actionable, they must be comprehensive and as real-time as possible.

Intel IT needed to increase velocity and provide greater value to sales and marketing with an integrated analytics platform that could do the following:

- Manage the growing diversity of sales and marketing data sources
- Overcome traditional data warehouse limitations
- Interconnect data models
- Improve data access
- Support accessibility across multiple business-intelligence (BI) front-end tools

Manage the Growing Diversity of Sales and Marketing Data Sources

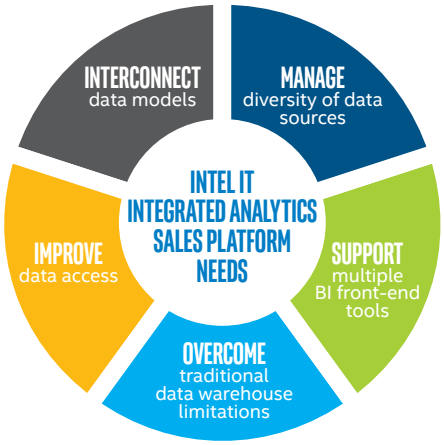
Sales and marketing organizations have a complex data source ecosystem. More than 20 gigabytes per day are ingested from thousands of internal sources—enterprise master data, departmental data marts, and spreadsheets—and external resources such as social media; blogs; third-party datasets; and data from distributors, resellers, and retailers.

Further, this data exists in multiple types, such as vendor APIs, log files, comma-separated value files, and flat files. We have more than 45 terabytes of data to manage for sales and marketing so far, and that amount is growing rapidly.

Overcome Traditional Data Warehouse Limitations

In traditional data warehousing, we needed to determine a data schema before ingesting the data. Data that was not included in this schema-first approach was never used, and its potential value was lost forever. Schemas changed too rapidly for internal and external sources to keep up with; the effort became too resource-intensive and time-consuming.

Inflexibility of data warehouses and schemas was slowing our velocity and limiting the value we provided to sales and marketing. With traditional BI data warehousing, we were spending 70 percent of our time defining schema prior to data ingestion and performing extract-transform-load (ETL) processes.



Interconnect Data Models

Data definitions, datasets, and data models were fragmented. Whenever sales and marketing requested a BI solution, we put together a project team to provide the solution. Because each project team was independent, their data models were not connected to other BI solutions and data models.

This lack of interconnected datasets weakened the value we provided to sales and marketing due to the limited information preventing a full view of the customer journey from marketing engagement to sales opportunity. If one Intel IT integrated analytics team built a data mart to monitor the campaign effectiveness for lead generation and another team was working on opportunity analysis for a sales pipeline, the data models created by these two separate teams would not be connected even though sales and marketing's strategy would benefit and find more value in a connected cross-sales pipeline approach.

For example, marketing decision makers could not confirm whether display advertisements or search-engine marketing keywords were reaching the personas being targeted by the sales pipeline. Even if we wanted to connect the models after the fact, we were unable to do so because data definitions were not aligned.

Improve Data Access

Business users in sales and marketing needed immediate access to the data. Unfortunately, it was taking us up to six weeks to add a single data attribute to a data model and make it available to business users. Additionally, it was taking up to six months to deploy BI capabilities.

Our lack of automation slowed data collection, cleansing, and alignment. Our lack of flexibility hindered access to data. Users who wanted to decrease the latency by using BI self-service still had to wait until a data model was created before they could perform their analysis, insert their results into presentations, and report their findings.

Support Accessibility Across Multiple BI Front-End Tools

We encourage business users to work with enterprise-level BI tools provided by Intel IT. However, when IT could not keep up with velocity of business demand, or when users did not like the user experience of a particular tool, then users often purchased other third-party BI tools.

We have learned how important it is to allow end users to work in their favorite tools when consuming data—and how often those favorite tools change—so we knew that we needed to design a BI solution that allows the users to use their tools of choice.

Solution

In 2014, Intel IT formed a team to start building an integrated analytics platform to give sales and marketing organizations a competitive advantage based on analytical insights and information collaboration. We are now building the Integrated Analytics Hub (IAH) platform with a data lake powered by Cloudera's distribution of Hadoop* (CDH). A data lake is a repository for large quantities of structured and unstructured data. We chose a data lake model because unlike relational data warehouses we can integrate large volumes of all data varieties in a single repository without the need to create schemas beforehand that define integration points between disparate data sets.

The repository can ingest data from various data streams at various rates and according to various data models and file formats. It also allows users to apply business schemas at any time, even while data trickles into the repository.

CDH is mature, open-source-based, big data processing software that provides scalable storage and distributed computing for our data lake.[1]

IAH increases the velocity of Intel IT's integrated analytics to keep up with sales and marketing's rapidly accelerating BI needs and connects multiple sources of data for end-to-end visibility by multiple users. When datasets are interconnected, analytics projects do not have to be independent of one another, and analysts and users can achieve economies of scale while achieving more actionable insights. As a foundation for advanced and predictive analytics, IAH helps users quickly formulate analytics questions and answers based on large volumes of data from multiple sources.

Solving Business Challenges with Velocity and Interconnected Data

The IAH project began by defining several technical capabilities to manage the volume, variety, and velocity of sales and marketing data.

To handle data growth and diversity, IAH has a data ingestion framework that supports a variety of data source types and protocols. After IAH ingests the data, the data lake powered by CDH allows us to work with data without needing to establish a schema first. This improves flexibility by allowing for customizable and extensible schema for metadata management, data profiling, distributed processing, and in-memory data manipulation. It also supports data quality rules that result in the ability to continuously evolve data models and resulting insights.

IAH improves data alignment and quality through data stewardship, which supports cleansing, conforming, and integrating disparate data sources. Sales and marketing data stewards can perform such functions

Data Lake Security and Privacy Controls

While the data lake allows for data agility and flexibility, it also provides tight data controls to address privacy and security issues, including the following:

- File-based extended access control attributes, both on the local system and Hadoop Distributed File System (HDFS), prevent unauthorized access.
- Lightweight Directory Access Protocol (LDAP) integration enables centralized identity management.
- LDAP integration enables fine-grained role-based access control and protected access to data through user-facing tools.
- Cloudera Sentry* enables fine-grained access control to data objects like Hive* tables, HDFS namespaces, and Apache Solr* collections.
- Kerberos* enables end-to-end authorized access to individual subsystems (user-facing or otherwise) within the data lake.
- Data encryption and key management technologies with Cloudera Navigator Encrypt* help us meet privacy and security compliance regulations.
- Transparent key management (in a physically separate repository of security certificates) and high availability of key subsystems makes the platform resistant to attack.
- Centralized operational management of the platform enables one-stop cataloging, auditing, and provenance capabilities for additional oversight of the data.

[1] For more information about how Intel IT uses Cloudera distribution of Hadoop, see "How Intel IT Successfully Migrated to Cloudera Apache Hadoop."

as mapping and aligning data to enterprise master data as well as business segment data, creating and defining line-of-business master data, defining the business glossary, and defining and applying data quality rules.

IAH expands data access by making data available in various states, including raw, cleansed, and conformed.

- **Raw data** is data in its original state with no transformations conducted during ingestion.
- **Cleansed data** is raw data that is processed by a master data management and data quality platform.
- **Conformed data** is cleansed data that has been modeled for data interconnection and aligned across other sources so that it is more accessible to users.

Users can access IAH's self-service capabilities to perform data modeling, merge data, and share data analysis results. Users can also explore, visualize, and process this data through their preferred BI front-end tools. IAH provides data in multiple formats to support the various BI tools within Intel's ecosystem.

The key to IAH increasing velocity and interconnecting data lies in the data lake powered by CDH, which is the foundation of the IAH architecture.

Integrated Analytics Hub Architecture

We chose CDH to implement the data lake for IAH because we wanted a data repository that stores data in its native format with no schema. CDH enables greater flexibility, automation, and faster access to data compared to using a data warehouse. Other components of the IAH architecture (illustrated in Figure 1) include a master data management platform; an in-memory platform; reporting tools; and the ability to share visualizations, reports, and dashboards.

IAH ingests data from more than 140,000 sources, including retail, media, marketing, customer relationship management (CRM), social, customer experience, and more. After the data is ingested, our goal is to make it available as raw data to data scientists within 24 hours compared to the six-month delay we experienced prior to the existence of IAH. Nothing has been done to the data at this point.

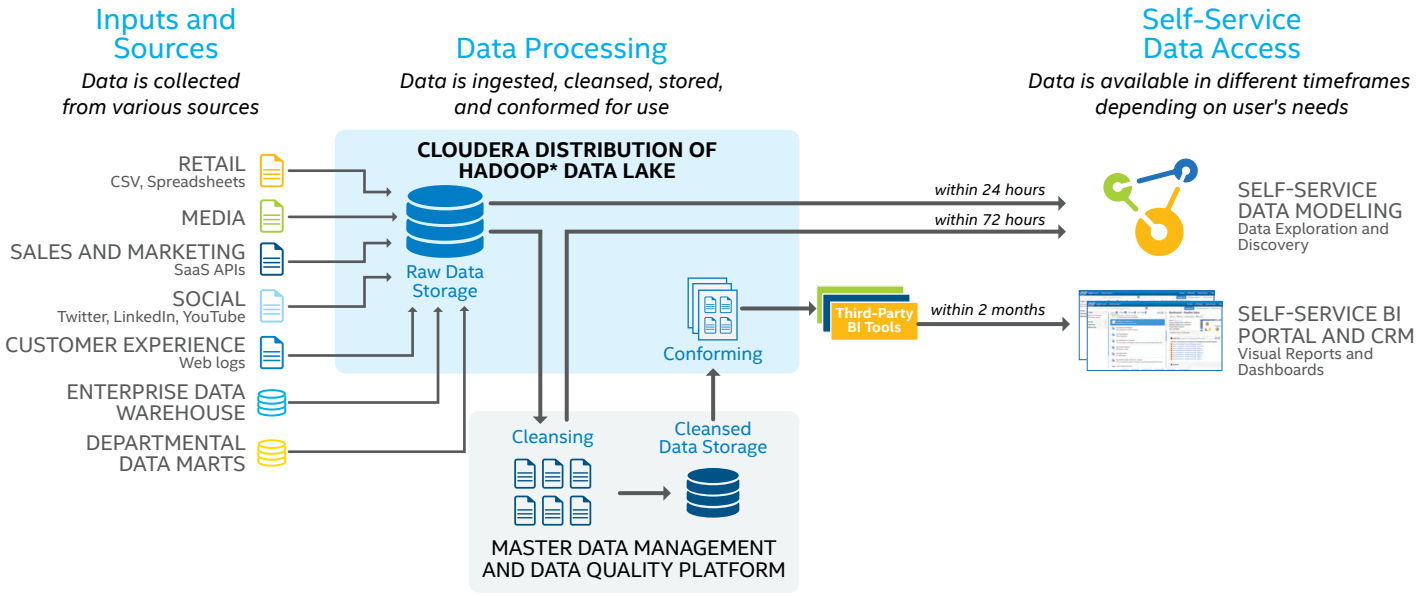


Figure 1. The Integrated Analytics Hub (IAH) architecture enables large volumes of raw, cleansed, and conformed data from more than 140,000 sources to co-exist in a data lake powered by Cloudera's distribution for Hadoop* (CDH). The goal is to make raw data available to data scientists within 24 hours.

50% FASTER

Time to align data definitions and model the data for complex queries, prior to implementing IAH.

When raw data is ingested into IAH, that data is retained in a nonpresumptuous, noninferred mode as well as an inferred mode. In the noninferred mode, no assumptions are made about the file type, data format, or field and record delimiters. All data is assumed to be binary and stored as such. In the inferred mode, raw data is analyzed for known file types, data formats, and field and record delimiters. All raw data is then rendered in a tabular structure to facilitate quick data exploration by the data analysts. Finally, all raw data is also presented in tables to accelerate tracking of the data provenance. On a use-case-by-use-case basis, we can apply the schema needed to answer individual business questions.

After raw data is ingested, a third-party master data management and data quality platform cleanses it to align with master data definitions, making it possible to interconnect the data. Our goal is to complete this process within 72 hours of when data ingestion begins. At this stage, BI analysts and business analysts can perform self-service data discovery.

Finally, most business users need the data to exist in prebuilt reports and dashboards. For simple queries such as sales operations data in one geographic area, we can automate this process. For more complex queries, Intel IT intervenes to align data definitions and model the data to meet business users' needs. While this process can take up to three months, that is still half as long as it took prior to implementing IAH.

Raw, cleansed, and conformed data types are exposed through BI front-end tools. Users can use the tools we provide—such as the Advanced Data Visualization Platform (Figure 2), which we built on an open source HTML5 and JavaScript framework to enable development of visual analytics applications—or other applications that they prefer.

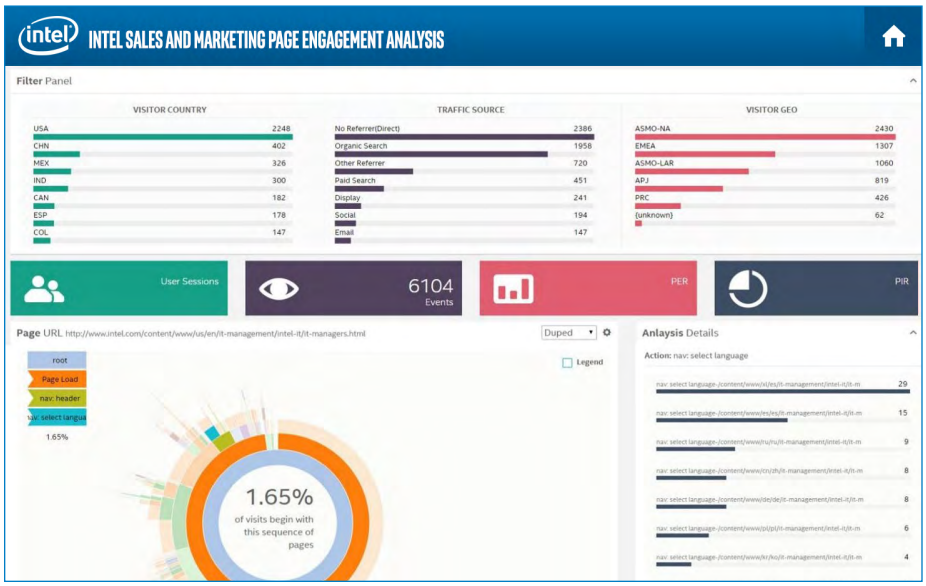


Figure 2. This page engagement analysis that we created for Intel sales and marketing organizations is an example of user-generated reports that are possible with the Advanced Data Visualization Platform.

Because we aim to enable users to consume data with any front-end BI tool, we expose the data in multiple formats. Users can create visualizations, reports, and dashboards and then publish and share them in the self-service MyBI Portal (Figure 3) or embed them within transactional applications systems, such as a CRM.

A New Paradigm for Self-Service BI

Automation capabilities in IAH reduce the need for manual Intel IT effort in the form of labor-intensive manual design and ETL processes. This has helped us keep up with sales and marketing's velocity and more than 140,000 data sources. It has also simplified self-service BI for users by making data readily available and easily understood. IAH provides data as relational database views, multidimensional cubes, in-memory tabular models, and raw data files, so users can access data with the BI tools of their choosing.

Self-service BI users rarely have to wait for IT to create data models. Most third-party BI tools connect directly to data sources, allow for immediate analysis and visualization within the tools, and then seamlessly embed results in presentation applications. Some presentation applications enable interactive experiences with the results, improving the user experience compared to using static images or screenshots. Users can also refresh data within the presentation application without needing to restart the process when new data is available. This saves time for self-service BI users.

More than 70 percent of sales and marketing analysts are self-servicing. IAH is one of the initiatives contributing to sales and marketing's year-over-year 40-percent increase in use of our self-service BI solutions and reports.[2]

Data stewards can now support BI analysts, business analysts, and end users without involving IT professionals and data scientists for every BI request. With IAH and other BI platforms, 84.4 percent of business analytics reports are generated without IT needing to intervene.

[2] For more information on self-service BI tools, read "Delivering Self-Service BI, Data Visualization, and Big Data Analytics."

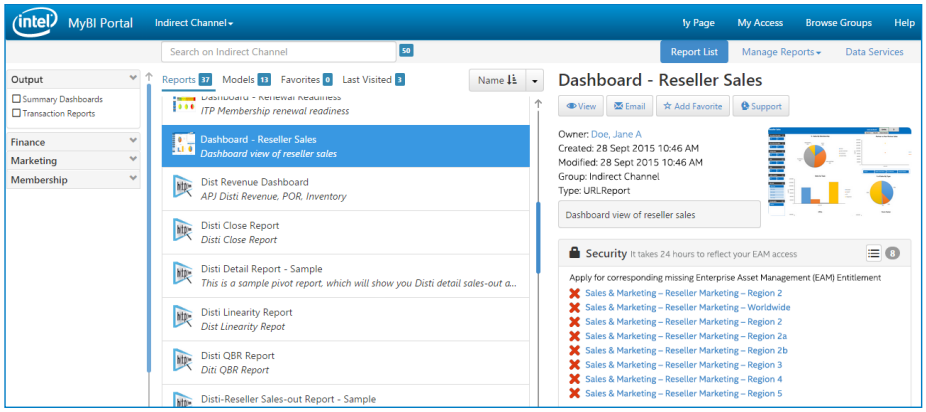


Figure 3. MyBI Portal is the self-service BI portal where users can share data visualizations, reports, and dashboards. The ability to share this information increases the velocity of insights and analytics, and removes manual IT intervention.

Preparing the IAH for Multiple Personas

Integrated Analytics Hub (IAH) helps multiple personas involved in the data analytics process. By increasing the velocity of making data available for analysis and by aligning master data definitions to make it easier to interconnect datasets, we made data analytics easier for some, made it more robust for others, and made it available to some new personas. IAH provides multiple ways to view the data, such as files, tables, datasets, and dataframes, to cater to each persona's preference.

- **IT professional.** IAH automates many of the tasks that this persona does manually. These experts in extract-transform-load operations, data modeling, data quality, and presentation toolsets support data access and self-service report creation.
- **Data scientist.** Working with raw, cleansed, and conformed data, this persona understands advanced statistics, data modeling, and programming. Data scientists create algorithms and predictive models to be consumed by end users.
- **Data steward.** This persona has extensive knowledge, but isn't an IT expert. They work with raw, cleansed, and conformed data and resolve ambiguities related to corporate vocabulary, manage data access, and ensure data consistency and integrity. They are the point of contact for data questions from BI analysts, business analysts, and end users.
- **BI analyst.** This business user can access cleansed data in less than 72 hours instead of waiting up to six months. BI analysts create and consume reports and dashboards frequently and are adept at merging multiple datasets, applying analytics techniques, and using BI tools.
- **Business analyst.** While skilled with data exploration tools, business analysts may rely on BI analysts and IT staff for data integration. This persona comsumes reports and dashboards to predict business outcomes and to help with decision making.
- **End user.** These are executives and managers who consume presentations, reports, and dashboards to inform their decision making.

Results

Architecture, design, and implementation of IAH are ongoing, but the platform is already in use and providing results. We have documented monetary savings, expedited BI capabilities, and exposed deeper insights with interconnected data.

For example, the marketing organization's Integrated Digital Media Analytics solution uses IAH to optimize marketing tactics across digital media channels. The CDH data lake allows users to consume, parse, integrate, and draw insights from a large volume of data in multiple formats. Insights from this analysis help marketing determine in which channel each piece of marketing content will be most effective. As a result, estimated quarterly savings on digital media expenditures is approximately USD 170,000.

BI capabilities are now available much faster than they were before we created IAH. Instead of waiting up to six months to access data, our goal is for users to have access to the following:

- Raw data within 24 hours.
- Cleansed data within less than 72 hours.
- Conformed data in approximately two months.

The ability to interconnect data in IAH also allows analysts to understand the direct impact of marketing automation efforts on sales. We have already identified a USD 576,000 final sale and determined how long it took for the lead to become an opportunity and, ultimately, a final sale. This level of insight informs strategic decision making on marketing and sales tactics.

Next Steps

As IAH matures, we continue to automate BI tasks and move analytics closer to decision making points. For example, we are also working to automate the development of data models in IAH to make it easier for users to gain insights from their data. A user initiates this process by uploading data into IAH and then selecting attributes within the dataset to trigger data alignment. IAH automatically generates an in-memory analytics model that users can query using their third-party BI tools. Self-service BI processes such as this require minimal technical skills, helping to democratize analytics for a broader audience.

Now that IAH is fully functional, we will focus on improving performance and streamlining architecture, including the following initiatives:

- Implement DevOps-like code management, automated unit testing, pair programming, peer review practices, and automated build and migration practices.
- Extend automated data profiling, data quality routing, and quality escalation processes to notify data custodians of syntactic and semantic violations in data.

- Create secure role-based access to protect data from end-to-end in time, location, and form.
- Improve software-as-a-service APIs around key ingestion, inference, and analytical functionality.

We strive to create a fully automated system that removes manual IT intervention from the entire data analytics process. Refinements to IAH's data ingestion processes and enhancements of IAH's metadata management system will improve efficiency and usability. We are also creating an automated monitoring solution for the data pipeline to provide transparency to business users on the status of the data being processed as well as enable our IAH support team to proactively respond to system or data issues.

Conclusion

Intel IT has created an integrated analytics platform to help the Intel sales and marketing organizations improve their competitive advantage. Using CDH to implement a data lake model, IAH standardizes dimensional data and connects datasets across sales and marketing. Increasing the automation of data analytics so that 84.4 percent of BI is self-service helps us keep up with sales and marketing's velocity, moving analytics closer to decision making points.

With IAH, we have accomplished the following:

- Reduced insight latency from months to days; our goal is reduce data insights and latency to 24 hours.
- Interconnected datasets so that analytics projects do not have to be independent of one another, achieving economies of scale and providing more actionable insights.
- Increased self-service BI by automating analytics tasks that traditionally need to be manually completed by an IT expert.
- Automated the alignment of enterprise and business segments across datasets while allowing our business partners to create, update, and own master data.

We continue to optimize IAH with greater automation and applications that take advantage of raw, cleansed, and conformed data that is ready to be explored.

For more information on Intel IT best practices, visit 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:

- 2015 Intel IT Annual Performance Report: Delivering Insights Worth Millions through Analytics
- Inside IT: Data Visualization podcast
- Big Data Analytics: Maximizing Marketing Insight through Data Analytics paper