PLANNING GUIDE

How to Move Forward with a Successful Deployment

Getting Started with Advanced Analytics

Table of Contents

Today's IT Landscape for Analytics

In IT circles, talk about big data has shifted from hype and initial skepticism to a more considered conversation focused on using new and emerging technologies to capitalize on data to create business value. No longer are people asking, "Is there value in big data?" Instead, they are asking, "How can I use data analytics solutions to create value for my organization?"

This change in tone reflects a new reality for all enterprises: maintaining the status quo is no longer an option. To compete effectively in a digitally driven world, organizations must put data analytics solutions to work to accelerate time to insight and gain a competitive advantage.

Given this compelling business case, it's no surprise that many companies are investing in analytics. These investments reflect the fact that from health care to retail, from banking to manufacturing, insights from massive amounts of data can enable organizations to make breakthrough discoveries, deliver better services, enrich the customer experience, and meet other business-driven goals.

Types of Analytics

To understand your organization's opportunities to capitalize on advanced analytics to drive insight, innovation, and competitive advantage, it helps to begin with the current state of the field. Analytics is a constantly evolving science that has changed dramatically over the years and continues to advance rapidly today.

Analytics now spans five categories: descriptive, diagnostic, predictive, prescriptive, and cognitive. These categories build on each other in a stepwise manner, as depicted in the analytics maturity model shown in Figure 1. Each step along this path moves the business toward an on-demand enterprise, where decision making becomes smarter and faster. It is important to understand where you are on the maturity path so you have a clear perspective on your current capabilities and where you want to go from there.

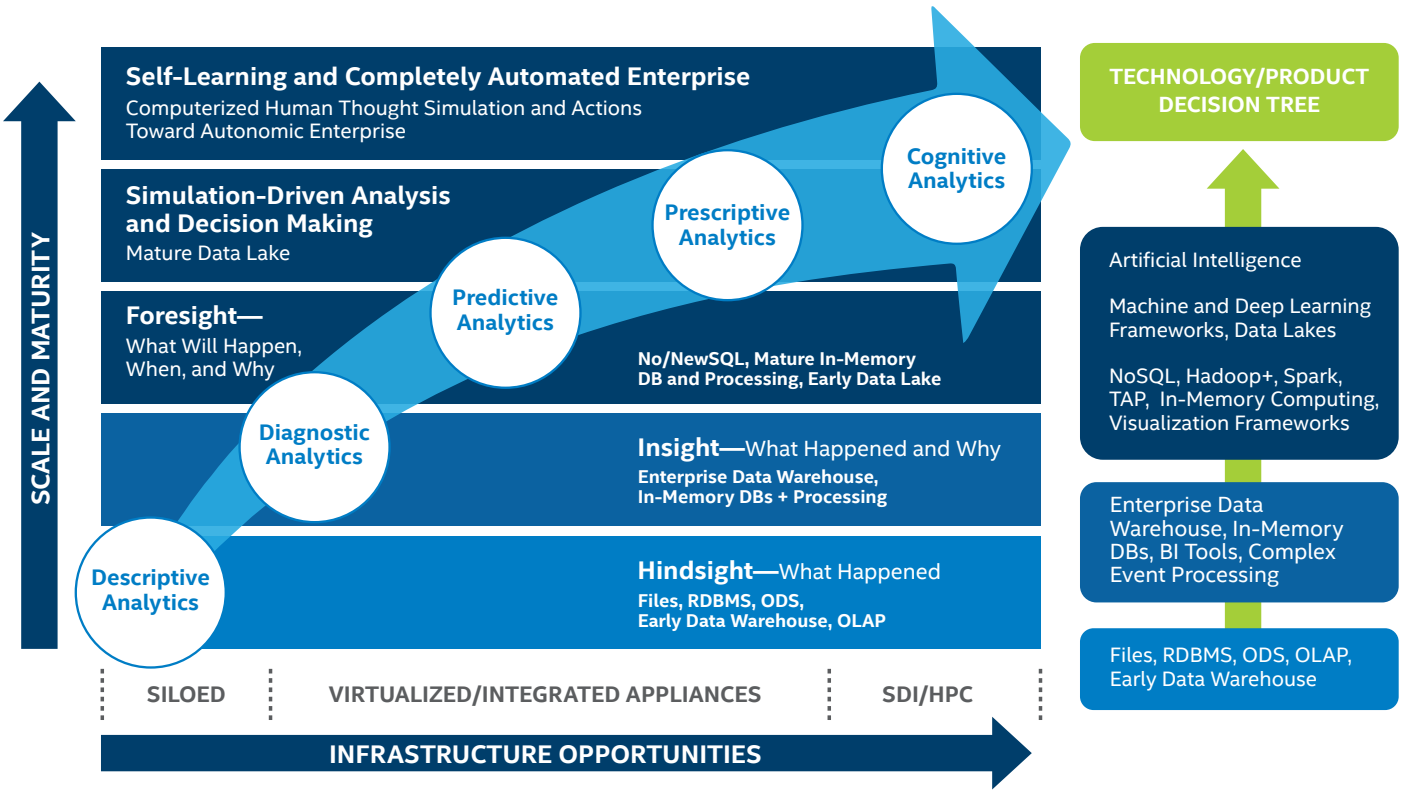

Figure 1. Advanced Analytics Maturity Path: Moving to Real-Time Enterprise
The analytics maturity curve moves from basic analytics that describe what happened in the past to cognitive analytics that automate decisions.

The categories of analytics fall into two groups: traditional and advanced.

Traditional Analytics

- Descriptive analytics answers questions about what happened in the past.
- Diagnostic analytics offers insight into why things happened.

Advanced Analytics

- Predictive analytics analyzes current and historical data to provide insight into what could happen in the future. It also allows us to infer what we would measure if we could, based on what we can measure. For example, we can't ask users if they are hungry, but we can infer that they are hungry if they are ordering a sandwich at lunchtime and they have a pattern of doing this in the past.
- Prescriptive analytics employs data techniques such as simulation and machine learning to suggest actions that a business could take to achieve the outcomes it desires.
- Cognitive analytics leverages artificial intelligence (AI) technologies, such as machine and deep learning, and high-performance data analytics to automate decisions using a human-like analysis, or augment human decisions through a partnership with smart machines.

Let's take a closer look at these stages in the analytics maturity path. Today, most companies have quite a bit of experience with the use of descriptive and diagnostic analytics to help them gain a better understanding of the business and the trends driving it. With this foundation in place, they are ready to progress to more advanced analytics that provide rich insights into where the business is going based on current trends and suggest strategic moves that the business can make to achieve the best outcomes.

As organizations evolve, they move from a focus on historical "what and why" questions to more forward-looking predictions and outcomes. This higher level of analytics leverages artificial intelligence (AI) techniques. Machine learning, a key enabling technology for AI, is a computational method that allows machines to act or think without being explicitly programmed to perform specific functions. Sets of algorithms, or mathematical models, "learn" from data. These models improve their performance on certain tasks based on experience, and then make predictions about new data.

Looking ahead further, companies are starting to think of cognitive analytics. This level of maturity involves a natural interaction-based man/machine collaboration, where human experience is augmented with smart machines to offer breakthrough insights for businesses.

IDC Forecast

By 2020, 40 percent of enterprises' net-new investments in analytics will be in predictive and prescriptive analytics.[1]

The Analytics Solutions Stack

The analytics solution stack encompasses four layers—infrastructure, data, analytics, and applications—with technologies that reach across all layers. From a functional standpoint, the technologies in the four layers complement each other and work together as a flexible big data platform that can also take advantage of an existing data management architecture to deliver both traditional and advanced analytics.

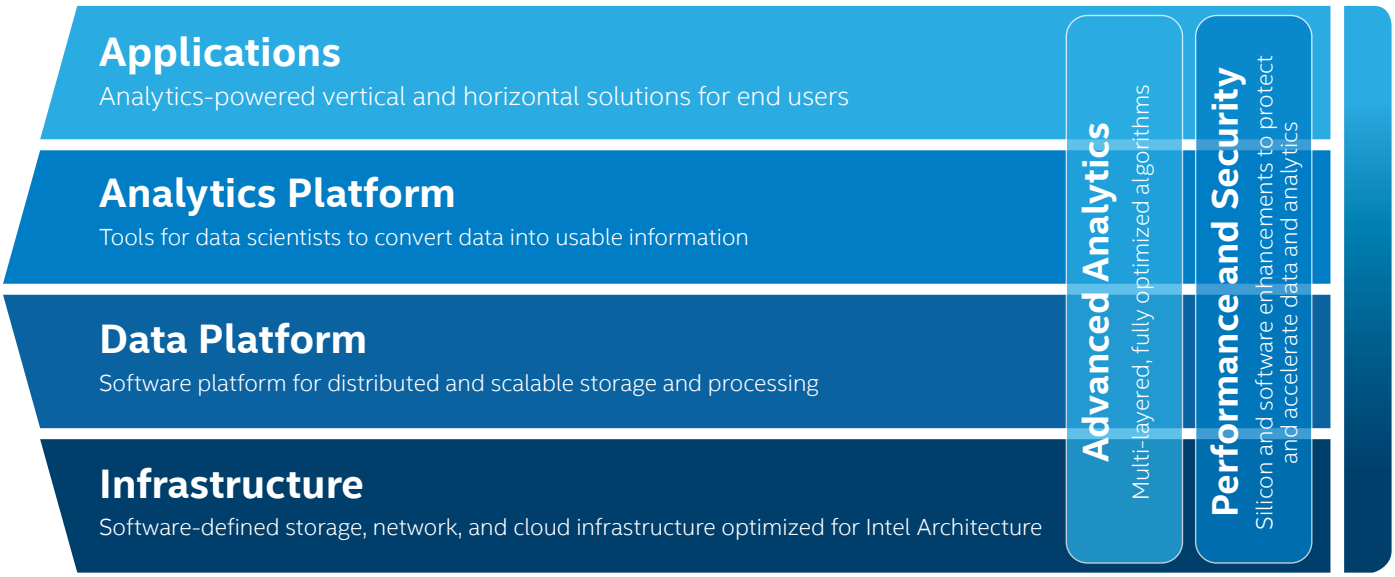

Figure 2. Analytics Solution
The analytics solution stack spans four layers: infrastructure, data, analytics, and applications.

The Four Layers

Infrastructure Layer
Whether the goal is to scale up to real-time analytics or scale out to include the big data sets in your analytics environment, or both, an analytics solution stack is built on a high-performance, scalable infrastructure layer. This technology foundation allows you to acquire, store, and protect data, and to run commercial and open source analytics solutions. Depending on the use case, the infrastructure layer typically powers some combination of an open source distributed processing framework, non-relational analytics databases, and analytics applications.

Data Layer
The data layer is where the data that the enterprise would perform its analytics on would reside. In the traditional analytics maturity level, the data layer consisted of primarily relational databases. Now, with the need to store and analyze streaming and unstructured data, this layer is supplemented by the Hadoop Distributed File System (HDFS), which can form the basis of your enterprise data hubs or data lakes. The data may also reside in NoSQL databases. Many enterprises will bring their ERP as well as other data, such as streaming data from the Internet of Things, into a real-time analytics environment, such as SAP HANA.

Analytics Layer
The analytics layer provides the building blocks for the analytic end-user applications. The analytics layer may have open source offerings such as Spark*, Storm*, suites from enterprise leaders such as Microsoft, SAP, SAS, Oracle, and IBM, as well as many smaller innovators.

Applications Layer
The top layer of the solution stack includes the out-of-the-box analytics applications across many industries that provide analytics to the various user types of the data.

Data Lakes

A data lake is a repository for large quantities of structured and unstructured data. It is more versatile than a data warehouse because it can ingest data from various streams at various rates and according to various data models and file formats. Gartner adds: "A data lake is a collection of storage instances of various data assets additional to the originating data sources. These assets are stored in a near-exact, or even exact, copy of the source format. The purpose of a data lake is to present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse)."[2]

NoSQL Databases

NoSQL databases provide high-performing, high-availability storage at Web scale. These nonrelational databases are useful for handling massive streams of data and flexible schema and data types with fast response times. NoSQL databases use a distributed and fault-tolerant architecture, which provides system dependability and scalability. Examples of NoSQL databases include Apache HBase, Apache Cassandra*, MarkLogic*, MongoDB*, and Apache CouchDB* software.

Other Technologies

From a functional standpoint, the technologies in the four layers complement each other and work together as a flexible big data platform that can also take advantage of existing data management systems to deliver traditional and advanced analytics. In addition, other technologies, such as artificial intelligence and performance and security solutions, run through all the layers of the solution stack, to accelerate advanced analytics insights and protect your data.

Artificial Intelligence (AI)
Data scientists, developers, and researchers are using tools on the path to AI, such as machine learning—leveraging algorithms or mathematical models that "learn" from data—to gain insights that were previously out of reach. The analytics solution stack facilitates and accelerates machine learning with multi-layered, fully optimized algorithms that enable computer models to generate insights from complex and unstructured data without explicit directives.

Deep learning is a branch of machine learning that is growing rapidly. It uses artificial neural networks that learn from experience—using deep graphs with multiple layers—to form models that can infer insights from new data. This approach is delivering breakthroughs in many human-like tasks, such as image recognition, speech recognition, and natural language processing. Combined, these approaches to machine intelligence are delivering business value today and transforming entire industries.

Performance and Security
Performance and security are paramount for a competitive, on-demand, and real-time enterprise. To that end, the analytics solution stack includes a wide range of silicon and software enhancements to protect and accelerate data and analytics.

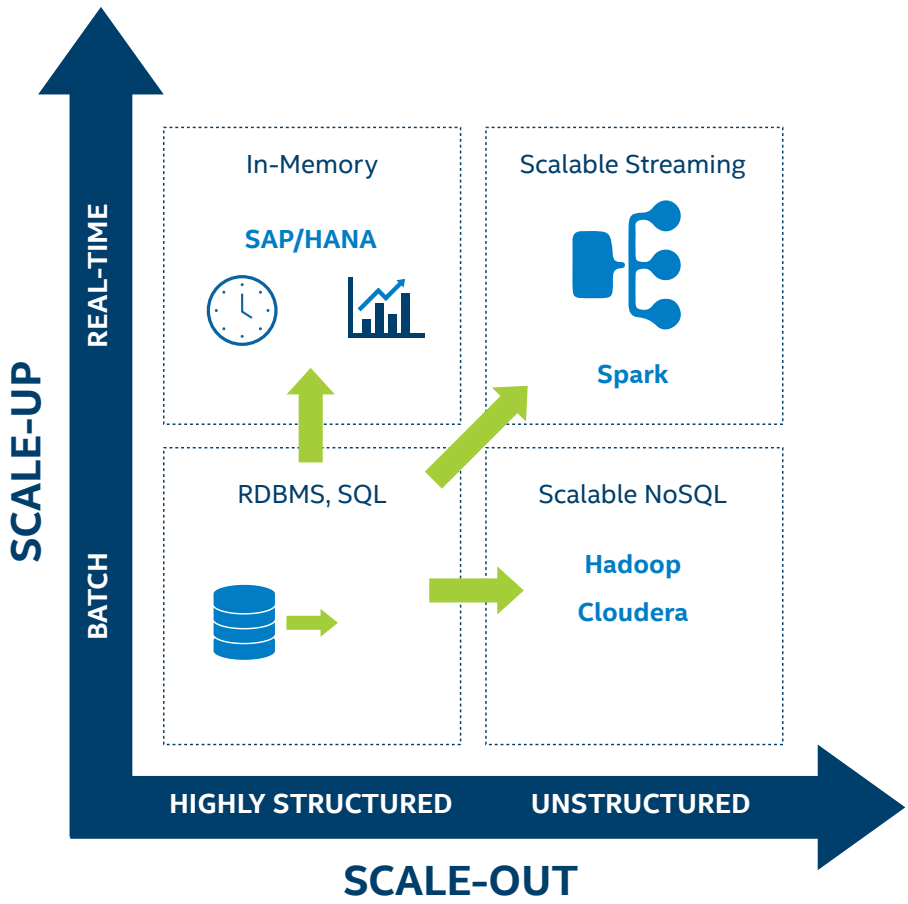

Figure 2. Analytics solutions can scale out or scale up to accommodate the variety, volume, and velocity of data.

NoSQL Databases

NoSQL databases provide high-performing, high-availability storage at Web scale. These nonrelational databases are useful for handling massive streams of data and flexible schema and data types with fast response times. NoSQL databases use a distributed and fault-tolerant architecture, which provides system dependability and scalability. Examples of NoSQL databases include Apache HBase, Apache Cassandra*, MarkLogic*, MongoDB*, and Apache CouchDB* software.

Getting Started with Analytics: Five Basic Steps

You now have a good understanding of the IT landscape for big data, its potential value to organizations, and the technologies that can help you get insights out of structured, semi-structured, and unstructured data resources. Plus, you have a good overview of the basics for getting the right infrastructure in place and running smoothly to support your big data initiatives.

You can get started with your big data analytics project by following the five basic steps we describe here.

Step 1: Focus on the business problems you are trying to solve.

Work with your business users to articulate the business opportunities.

- Identify and collaborate with business users—including analytics, data, and compliance officers; data scientists and stewards; citizen data scientists; and developers—to find the best business opportunities for big data analytics in your organization.
- Consider an existing business problem—especially one that is difficult, expensive, or impossible to accomplish with your current data sources and analytics systems. Or consider a problem that has never been addressed before because the data sources are new and unstructured.
- Prioritize your opportunity list and select a project with a discernible return on investment. To determine the best project, consider your answers to these questions:
  - What am I trying to accomplish?
  - Does this project align with strategic business goals?
  - Can I get management support for the project?
  - Does big data analytics hold a unique promise for insight over more traditional analytics?
  - What actions can I take based on the results of my project?
  - What is the potential return on investment to my business?
  - Can I deliver this project with a 6- to 12-month time to value?
  - Is the data that I need available? What do I own? What do I need to buy?
  - Is the data collected in real time, or is it historical data?

Collaborate with business leadership on a big data strategy and approach. Develop:

- The business case for analytics – Define how analytics drive value for your business. Identify the key business challenges analytics solutions will address.
- Short-, mid-, and long-term objectives – Outline the key phases to achieving your analytics goals.

Step 2: Understand how analytics will impact your culture and operations.

Develop a closer understanding of data analytics solutions.

- Talk with your peers in IT and the business.
- Take advantage of Intel resources for analytics to get up to speed on the technologies.
- Understand vendor offerings.
- Leverage tutorials and examine user documentation, such as that offered by the Apache Hadoop project.

Evaluate your infrastructure and operational needs. Consider:

- The current and future state of your IT infrastructure – Can your data center support the big data platform? Assess your current data center technology and describe, if necessary, your plan to upgrade computing, storage, and networking resources.
- Data sources and data quality – What are the primary sources of data internally? What additional data might you purchase? How will you ensure quality?
- Analytics platform and tools – What platform will you use to build your solution? What software and tools are needed to achieve your purpose?
- Metrics for measuring success – How will you measure system performance? Base your success on how many jobs are submitted, parallel processed, and completed efficiently.

Step 3: Identify and cultivate the skills you need.

Understand and plan for the skills required in both business and IT.

- Identify the skills you need to successfully accomplish your analytics initiative.
- Determine if the needed resources are in-house.
- Determine if you can build skills from within the company.
- Determine where your analytics professionals will reside within the business or IT organization.
- Hire new talent or outsource certain functions as needed.

Step 4: Consider your technology requirements.

Identify the gaps between current- and future-state capabilities.

- What additional data quality requirements will you have for collecting, cleansing, and aggregating data into usable formats?
- What data governance policies will need to be in place for classifying data; defining its relevance; and storing, analyzing, and accessing it?
- What infrastructure capabilities will need to be in place to ensure scalability, low latency, and performance, including computing, storage, and network capabilities?
- Do you need to add specialized components like a NoSQL database for low-latency lookups on large volumes of heterogeneous data?
- If you plan to process a steady stream of real-time data, what additional infrastructure and memory capabilities will you need? Will you require an MPP in-memory analytics appliance? A CEP solution?
- Identify the analytical queries and algorithms required to generate the desired outputs.
- If you are considering cloud computing for your delivery model, what type of cloud environment will you use? Private, hybrid, public?
- How will data be presented to users? Findings need to be delivered in an easy-to-understand way to a variety of business users, from senior executives to information professionals.

A Business-IT Collaboration

With so much at stake for the business, analytics initiatives can't happen in a vacuum. IT must forge a strong partnership with business leaders to identify big data opportunities and move forward with needed support. Analytics initiatives also require new business, technical, and analytical skill sets to help model complex business problems and discover insights, integrate systems, build out massive databases, and administer distributed software frameworks.

Implementing advanced analytics in your business means your data now takes a prominent place at the business table at the highest levels, and you need to have proper representation for it and your processes. You need to rethink the roles in your organization around your data, addressing new ways of managing, interpreting, and analyzing vast amounts and types of data.

Some of these roles are fairly well established, while others are fairly new. The newer roles include chief data scientist, chief analytics officer, and citizen data scientist:

- Chief data scientists combine expertise in computer science, mathematics, statistical modeling, and analytics to extract actionable insights from big data.
- Chief analytics officers oversee an organization's use of analytics tools to gain insights and value from data.
- Citizen data scientists use analytics tools to generate insights from data but hold positions that are not focused on big data or analytics. They could come from anywhere in the organization.

In another notable trend, data analytics teams increasingly work closely with people in different lines of business to help them exploit data to drive the business forward.

Step 5: Implement your data solution.

Develop use cases for your project.

- Identify the use cases required to carry out your project.
- Map out data flows to help define your required analytics capabilities.
- Decide what data to include and what to leave out. Identify only the strategic data that will lead to meaningful insight.
- Determine how data interrelates and the complexity of the business rules.
- Consider whether you need to support advanced analytics, such as interactive queries or predictive analytics, or support real-time data streams.

Develop a test environment for a production version.

- Adapt reference architectures to your enterprise. Intel is working with leading partners to develop reference architectures that can help as part of the Intel Cloud Builders program around big data use cases.
- Define the presentation layer, analytics application layer, data warehousing, and, if applicable, private- or public-based cloud data management.
- Determine the tools users require to present results in a meaningful way. User adoption of tools will significantly influence the overall success of your project.

Consideration for Moving Forward: A Decision Tree

To help identify the right analytics solution for your organization, you can follow a decision-tree process similar to that shown in Figure 4. Your answers to the questions in the decision tree can help your team better understand your specific requirements and technology needs.

In this decision tree, technology and analytics factors are depicted in blue shading. Organizational factors are depicted in green shading. The flow begins with a clear use case that is consistently understood by LoB and technology stakeholders. This use case should be aligned to the organization's place on the analytics maturity curve, as shown above in Figure 1. For example, if an organization were at the diagnostic level of the maturity model, and focused on getting its data through daily batch processing, it would not be ready to cater to a use case that required real-time predictions.

The targeted use case should be examined to determine if it requires real-time analytics or if batch will suffice. Also, the data required for the use case should be defined—is it structured data from an internal ERP system or external data that may be unstructured? The flow chart shown in Figure 4 also provides high-level guidance for how organizational and technology factors would influence the choice of technologies and skills that an organization would require.

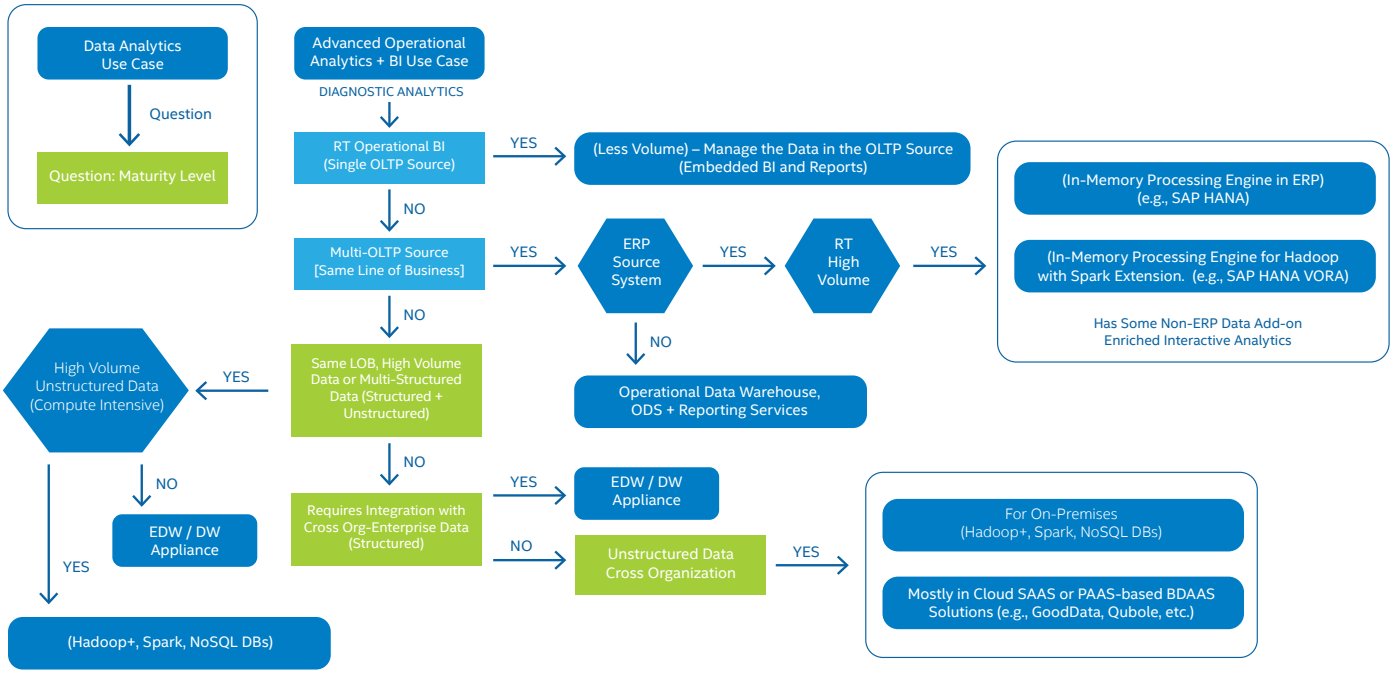

Figure 4. A decision tree for analytics.

How Intel Can Help with Your Analytics Journey

Intel is involved at nearly every level of the data analytics solution stack, not only as the provider of big data infrastructure but also as a key contributor to open source projects and an active supporter of ISV optimization leads to innovations in analytics applications, interoperability, and better performance on Intel® architecture. Together, Intel and its ecosystem partners provide the technologies for a comprehensive advanced analytics solution stack that covers the infrastructure, data, analytics, and applications layers.

In addition, Intel not only provides the foundation for the solution stack through leading compute, storage, and networking but also provides security and performance-acceleration solutions along with software libraries and accelerators that are used by the layers that sit on the infrastructure foundation.

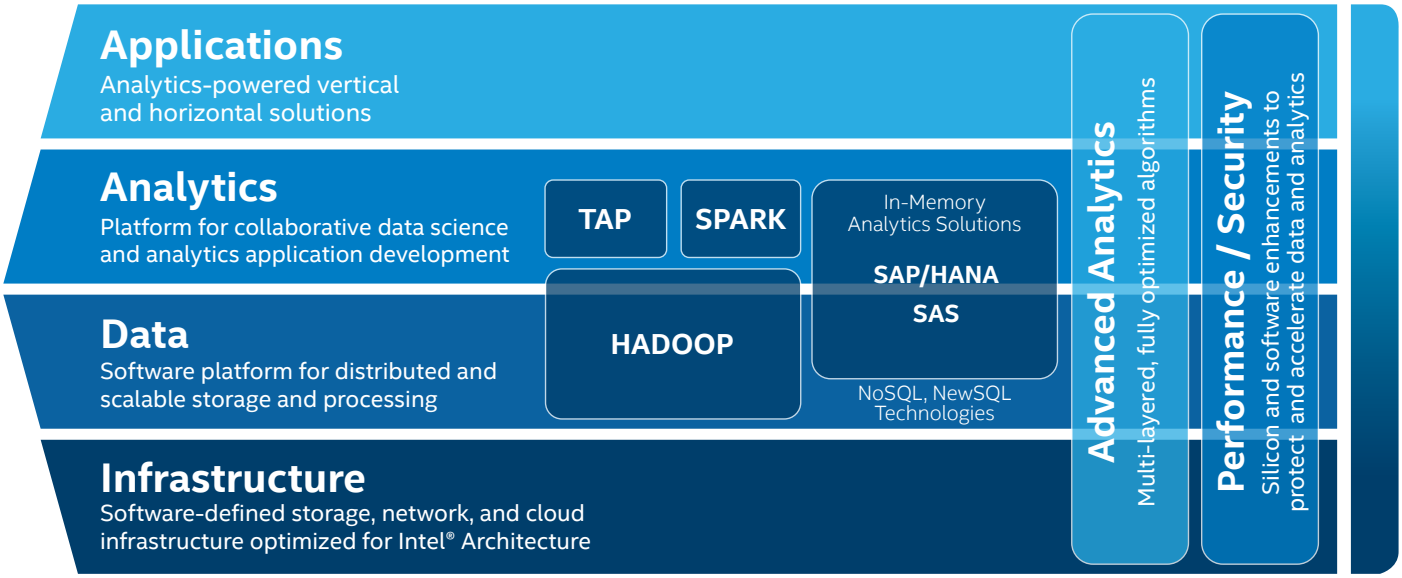

Figure 5. Analytics Solution Stack
Intel technologies and contributions span the entire analytics solutions stack.

Here are a few examples of Intel® products and technologies that provide the foundation for analytics across data center infrastructure.

Compute
Intel's rich portfolio of workload-optimized processors covers the full range of analytics needs:

- The Intel® Xeon® processor E5 family offers an ideal mix of performance, efficiency, and features for scale-out analytics deployments and machine learning execution, also known as inference.
- The Intel® Xeon® processor E7 family offers expanded memory capacity to enable in-memory analytics and scale-up architectures.
- The Intel® Xeon Phi™ product family excels at tasks like simulations and machine learning algorithm development.

Storage
While data storage capacity is growing, floor space, power, and budgets are not. Breakthroughs in storage media, such as solid state drives, are redefining the cost, capacity, and latency of storage. Storage solutions designed with Intel Xeon processors provide fast and dependable storage, specifically designed and optimized for software-defined infrastructure.

Today, Intel is ramping up next-generation storage solutions with offerings like Intel® Solid State Drives—designed for seamless performance and enhanced capabilities with Intel CPUs, chipsets, firmware, software, and drivers—and new 3D XPoint™ Technology, which enables blazing-fast speeds, lower prices, and enhanced endurance.

Networking
Delays in moving your data from your data lake or warehouse to your compute infrastructure can be costly in real-time operations. To help your organization avoid these delays, Intel provides one of the fastest fabrics available—the Intel® Omni-Path Architecture. Based on Remote Direct Memory Access (RDMA), Intel Omni-Path Architecture is the next-generation fabric for high-performance data analytics and high-performance computing. In addition, Intel is a world leader in providing Ethernet adapters and converged adapters for the data center. To date, Intel has shipped more than a billion Ethernet ports.

High-Performance Computing
As more enterprises turn to analytics to tackle increasingly complex problems, and want results in near real time, there is a growing interest in HPC. Intel can help your organization cut through the complexity of deploying HPC clusters with the Intel® Scalable System Framework (Intel® SSF), which allows organizations to accelerate innovation with breakthrough performance to run analytics workloads on a common infrastructure.

Security
Intel hardware and software security tools help block known threats, identify compromises, and expedite remediation to securely enable a safe and connected world. Intel includes security features like Intel® Advanced Encryption Standards New Instructions (Intel® AES-NI) to boost cryptography performance to allow for more data to be kept secure while it's stored, while it's being transmitted on the network, and while it's being analyzed.

Software Libraries and Accelerators
For most of these tasks, computational speed is a key ingredient for success. Intel helps you get there with innovative software libraries and accelerators that optimize frameworks for higher performance on Intel architecture.

- The Intel® Data Analytics Acceleration Library (Intel® DAAL) is designed to help software developers reduce the time it takes to develop their applications and deliver them with improved performance. Intel DAAL helps applications make better predictions—faster—and analyze larger data sets with the available compute resources at hand.
- The Intel® Math Kernel Library (Intel® MKL), meanwhile, accelerates math processing routines that increase application performance and reduce development time. Intel MKL includes highly vectorized and threaded Linear Algebra, Fast Fourier Transforms (FFT), Vector Math, and Statistics functions. The easiest way to take advantage of all of that processing power is to use a carefully optimized computing math library, because even the best compiler can't compete with the level of performance possible from a hand-optimized library.
- Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) supports CPU-optimized artificial intelligence operations through the inclusion of deep neural network (DNN) performance primitives for DNN applications. These open source software libraries help users develop and run deep learning applications optimized for Intel architecture.

Ultimately, Intel's open source software is making it easier to access AI, from developing your own AI framework using Intel Math Kernel Library for Deep Neural Networks, to building advanced neural networks, to bridging the gap between big data and cloud applications.

The Intel IT Analytics Platform

In addition to working with ecosystem partners to help customers build analytics platforms, Intel is putting its platform-development expertise to work in-house across a wide range of analytics use case. In 2014 alone, Intel IT's use of BI and analytics tools increased Intel revenue by USD 351 million.[3]

Consider this example of the way Intel IT is capitalizing on data analytics. Intel IT is building an analytics platform that integrates and connects business intelligence (BI) data using a data lake model to reduce insight latency to 24 hours. This Integrated Analytics Hub (IAH) helps the Intel sales and marketing organizations make smarter data-driven decisions.

Among other benefits:

- The data lake model provides data scientists, analysts, data stewards, and end users with faster and more flexible access to large volumes of data that is available in multiple formats and in three states—raw, cleansed, and conformed—for various levels of analysis.
- All users can easily engage in self-service BI and use BI front-end tools of choice for analysis of data at any of the various three states.
- The ability to interconnect data sets and share visualizations, reports, and dashboards on the self-service BI portal increases velocity and removes manual IT intervention.

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

For the full story, read the white paper: How Intel IT's Integrated Analytics Platform Helps Sales and Marketing.

Additional Analytics Resources

In addition to the resources already cited in this paper, check the following for further interesting content.

Intel Program and Product Web Sites

- Advanced Analytics with Intel
- Analytics blogs and insights
- Machine Learning
- Artificial Intelligence
- Intel Xeon processor E5 family
- Intel Xeon processor E7 family
- Intel Xeon Phi product family
- Nervana Systems
- Saffron Technology

Intel IT Web Site
Sharing IT Best Practices
Access white papers, reports, and other resources focused on best practices in IT, including those for analytics, big data, and the Internet of Things.

Intel IT White Papers
Optimizing Intel's Supply Chain with an In-Memory Data Platform
Learn how a new in-memory data platform will transform Intel's supply chain by providing real-time predictive business analytics to enable better, faster decision making.

Big Data: Securing Intel IT's Apache Hadoop* Platform
Explore Intel IT's use of Apache Sentry* and Cloudera Navigator* to secure an Apache Hadoop* platform at the perimeter, access, visibility, and data levels.

How Intel IT Successfully Migrated to Cloudera Apache Hadoop*
Get a close-up look at Intel IT's migration to Apache Hadoop software and six best practices developed by the Intel IT migration team.

Improving Manufacturing with Advanced Data Analytics
See how Intel IT is using sensors, the Internet of Things, and data analysis to improve product quality, reduce capital costs, and speed time to market.

Using Big Data in Manufacturing at Intel's Smart Factories
Learn how Intel's smart factories use edge computing and the Internet of Things to enable automated control systems with real-time data.

Intel Ecosystem Partner Solution Sites
Intel works actively with its ecosystem partners to optimize solutions for interoperability and better performance on Intel architecture. For examples of the work, see these partner pages.

Cloudera
Partnering with Cloudera, Intel has helped deliver enterprise-grade innovations to the Hadoop* framework in security, performance, management, and governance.

SAS
With SAS Analytics* 9.4 optimized for Intel® Xeon® E7 processors, organizations can see performance increases in predictive analytics workloads on one new powerful system.

SAP
Through the joint optimization of SAP* HANA and Intel® Xeon® E7 processors with TSX, organizations get performance improvements for real-time transactions and analytics.

[1] IDC, Big Data Forecast, November 2015.

[2] Gartner IT Glossary: Data Lake.

[3] Achieving Intel Transformation through IT Innovation: 2014–2015 Intel IT Business Review – Annual Edition.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/big-data/sap-hana-real-time-analytics-solution-brief.html.