FROM DATA TO INSIGHTS

Optimize the four stages of your data pipeline to make your business analytics-driven
The volume and variety of the data your organization holds or accesses today is much greater than it was just a few years ago. Next year it will be greater still. As your data resources grow, so do the opportunities to draw deeper, more powerful, insights about your customers, competitive landscape and internal processes. However, in many cases, ageing IT architectures and methodologies are hampering the ability to seize these opportunities. In fact, according to McKinsey, just 8 percent of companies today say they have been able to productionalize and scale their analytics programs¹.

This guide explores and provides guidance on the three key steps to becoming analytics-driven:

- **Building an organizational foundation**
- **Mapping the data pipeline**
- **Productionalizing analytics**
ORGANIZATIONAL FOUNDATION:
RALLY THE NEED AND IDENTIFY BUSINESS VALUE

Start by working with the right people at all levels of the business to create a solid foundation for your analytics-driven organization. Gain strong and unified commitment from management — from the C-Suite down — by helping them understand the value of your proposed analytics projects to them and to the business as a whole.

Make sure as well that you have the human resources you need at an operational level. Invest in building the expertise you need, either through hiring or training, to deliver on your plans. Create cross-functional teams that include stakeholders from business, technical and analytics perspectives to ensure you can take a collaborative and agile approach.

Once you have your team in place, work together to identify and prioritize your analytics projects (see figure 1). This is a crucial stage, which deserves concentrated discussion. As Albert Einstein said: “If I had an hour to solve a problem I’d spend 55 minutes thinking about the problem and five minutes thinking about solutions.” Brainstorm business problems together and what it would take to solve them — including the questions to ask, the data needed to answer them, and the processes or resources needed to implement the insights.

While there may be great enthusiasm to deploy analytics in a range of ways all at once, don’t be tempted to boil the ocean. Start with your top use cases (no more than five), get some successes under your belt, and then build momentum as you gain experience. It’s critical in these early cases that you can quantify the impact on or value for the business to secure ongoing management support.

Figure 1: Work with your team to prioritize analytics use cases
Also important is the ability to embed your analytics use cases into business processes to make implementation as seamless as possible. If it’s going to require a heavy lift in terms of process or cultural change, pick something else.

A strong foundation for analytics involves an end-to-end data strategy as much as a strong team. Make sure those involved understand the needs and characteristics of your data today, and for future use cases. How will you handle the need to migrate and/or scale your data as your needs change? Implement a master data model across key domains (such as customer, product and locations) to create consistency, and be clear about ownership and accountability for security, governance, quality and maintenance of your data.

Only once these issues have been fully thought through and addressed can you begin to integrate, embed and productionalize analytics into your decision-making processes.

Figure 2 showcases the various components needed to establish an effective organizational foundation to develop analytic work. Use this analytics and AI organizational evaluation workbook, which contains critical thinking questions around each component for the organization, to evaluate how ready your organization is to support its analytics and AI journey.

**Figure 2:** The organizational foundation for analytic work

**Define the Problem**
A business problem that, with the use of data, will provide more insight to drive the right business decisions or provide business efficiency on how work is to done, which is sponsored by management direction.

**Problem Solving Process**
Team breaks down the defined business problem into workable steps to translate the right data to achieve results.

**Right Software & Hardware Infrastructure**
Organization secures infrastructure that supports data processing in a timely manner.

**Right Team of Experts**
A team of management sponsors, data scientists, data engineers, solution architects, and domain experts identifies the right data and works to translate the data to achieve results.

**Culture and Resources**
Organization embrace data insights, sponsors properly resourced teams, and prioritizes analytic development work.

**Continuous Improvement Practices**
Team embraces fail-fast continuous improvement practices to evaluate their success in translating data to achieve results.

**Right Data**
Team understand and obtains the right data that explains the business problem to achieve results.
DATA PIPELINE: THE MECHANICS OF DATA STRATEGY, HANDLING AND PROCESSING DATA

It’s often possible to make a start with your existing infrastructure and tools, and begin a discussion on forming strategic investments in new resources and iterate once use cases have been defined and business value demonstrated. In order to do this, it’s important to consider every stage of your data lifecycle – ingest, store, process and analyze – and make adjustments or optimizations where they will have the most impact. This guide is designed to help you think through where those areas might be in your organization.

IDENTIFYING THE ROADBLOCKS TO ANALYTICS READINESS

Data continues to pour into organizations, in all shapes and sizes. IT leaders must wrangle these unwieldy resources to drive revenue, automate processes, increase efficiencies, and create new products or services. Being unable to do so risks missed opportunities and bottlenecks in decision making. The pain points facing most IT departments can be broken down into two main areas.

Firstly, poor-quality or incomplete data is always a risk here, but the time and effort needed to clean, check and prepare even good-quality data can be excessive for small and busy IT teams. Secondly, data often resides in siloes around the organization, with no easy way to view or analyze it as a whole. Decisions on where it resides and how best to use it are further controlled by governance and compliance requirements, and by the increasing complexity of environments that span on-premises, private and public cloud. It may make sense to run one workload in the public cloud one day (for example, training an artificial intelligence (AI) algorithm) and then move it to on-premises environments for the next phase (iteration). The shift must be seamless, without compromising access to the necessary data.

OPTIMIZE EACH STAGE OF THE DATA LIFECYCLE

Creating the most valuable business insights requires an end-to-end analytics strategy across the data pipeline or lifecycle (see figure 3). At every stage of this data continuum, you must address the inherent challenges and focus on the desired outcome of each phase, while also addressing overarching concerns such as security and governance.

The data lifecycle runs from the moment of data creation, through ingestion to staging to analytics, visualization and eventually archival. It is not always a linear process nor is it one dimensional – there are usually many subroutines and data morphing along this path. There can be overlap as well with, for example storage playing an important role at the ingest and process stages. However, it can be a useful way of thinking about your overall data strategy.

<table>
<thead>
<tr>
<th>Ingest</th>
<th>Prepare</th>
<th>Analyze</th>
<th>Act</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data creation</td>
<td>Data transmission</td>
<td>Data ingestion</td>
<td>Data integration</td>
</tr>
<tr>
<td>Data staging</td>
<td>Data clean, normalise</td>
<td>Models and experiment</td>
<td>Tuning and distillation</td>
</tr>
<tr>
<td>Visualize and deploy</td>
<td>Archival</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: The stages of the data lifecycle, or data pipeline
The first stage is to pull in your raw data. Where it comes from should be determined in collaboration with line of business (LoB) colleagues, who can help identify what data is needed, and (just as importantly) what is irrelevant for generating the required insights. Ingestion usually involves connecting to various data sources, extracting the data, and detecting what’s changed before moving it from its source into a system where it can be stored and analyzed (the target).

There are a number of approaches you can take to collect raw data, based on the data’s size, source, structure and latency. For example, data from application events, such as log files or user events, is typically collected in a push model, where the app calls an application programming interface (API) to send the data to storage. Collecting and analyzing this event-driven data can reveal user trends and provide valuable business insights.

Meanwhile, streaming data is made up of a continuous flow of small, asynchronous messages, which are delivered without expecting a reply. This type of data has two common usages:

- Telemetry - collecting data from geographically dispersed devices like edge sensors. These devices individually collect relatively small amounts of data, but at scale they can generate huge data volumes that require complex analysis and are often used as input for machine learning tasks.

- User events and analytics – collecting data from user devices about use cases and behavior patterns, such as when an app is opened, or when it crashes. The aggregate of this data across all the mobile devices where the app is installed can provide valuable information about usage, metrics and code quality.

We are currently seeing a trend towards more real-time analytics of streaming data (for example cybersecurity monitoring), which requires immediate analysis and insights. This demands greater distributed compute and memory (in virtual machines (VMs) and/or at the edge) to enable real-time action. In particular it needs high compute performance at low power. Intel® CPUs that are optimized for low-power, high-performance edge workloads can help meet these requirements.

In comparison, a manufacturing supply chain manager wishing to gain insights into trends across his pipeline, would need to focus on bringing together data streams from different siloes (supply, demand, inventory etc) and run analytics to identify trends after the fact. This would rely more on batch data, which tends to consist of large numbers of files that are transferred and stored in bulk, or in relational or NoSQL* databases. The data can be located on-premises or on other cloud platforms, and ingesting it requires high aggregate bandwidth between the sources and the target.

As with ingestion at the edge, high-performance CPUs are essential in ingesting batch data, with the added requirement of massive scalability as data volumes grow. The latest data center technologies from Intel, including Intel® Xeon® Scalable processors, Intel® Optane™ DC persistent memory and Intel® Optane™ DC Solid-State Drives, are architected from the ground up to meet these requirements.
up for this type of challenge, with support for a new, bigger tier of memory and extreme performance for fast-growing analytics and AI workloads.

Your compute, networking and storage needs will change depending on the source, type and frequency of your data ingestion. For example, at the point of ingestion, you may choose to use the ‘extract, transform, load’ (ETL) model for structured, heterogeneous data, putting it straight into storage for later analysis. Meanwhile, a real-time analytics use case based on streaming data would be better suited to an ‘extract, load, transform’ (ELT) approach, which loads and writes the data, then tweaks it to get the most out of the stored information. The ‘ELT’ model puts the burden of cleaning the data more into the third stage of the data lifecycle (process), which we’ll explore below, but it’s important to consider which model will work best for each use case from the outset.

**AVOID COMMON MISTAKES WHEN SETTING YOUR INGESTION STRATEGY**

* Don’t assume ingestion is a quick process. The ingestion stage can be complex and can take time. As data volumes and variety grow, so will the time data scientists must spend extracting, loading and transforming it before it can be analyzed. Take this into account and allow time for it in your processes.
* Keep your data clean. It’s important to make sure your data stays clean over time. Set governance in place to ensure this; the nearer to the time of capture or ingestion the better.
The key consideration from a storage perspective is data access – ensuring that the applications and programs running analytics get the data they need when and where they need it. This means making sure compute performance, memory capacity and network bandwidth are aligned to keep the data the application needs as close and accessible as possible. At the end of this stage, the aim is to have your data in a format that enables it to be easily accessed and queried. For example, you might clean training data to use in fine-tuning a machine learning model, or you could transform raw data so that it can be ingested into data warehousing tools or Hadoop for analytics.

When considering storage, be led by the needs of your target workload. Object stores are used for storing structured and unstructured data like logs, database backups, files, images, and videos. For this type of storage requirement, consider QLC NAND devices for cost-effective, high-volume storage, and solid-state drives (SSDs) for caching.

For databases, a SQL relational database management system (RDBMS) (e.g. SAP HANA, Microsoft SQL Server, Oracle DB, MySQL) is generally suited to online transaction processing (OLTP) workloads with structured data that require ACID compliance, like financial transactions, customer order tracking or checking user credentials. On the other hand, NoSQL databases (e.g. Redis, Cassandra, HBase) tend to support terabyte- to petabyte-scale OLTP workloads that support real-time and/or large-scale analytical workloads, such as real-time app data, Internet of things (IoT) sensor data or advanced analytical and machine learning tasks. These accelerated workloads demand CPUs for faster processing, and 3D NAND or persistent memory to support real-time operations.

Delivered in conjunction with the 2nd generation Intel Xeon Scalable processor, Intel Optane DC persistent memory is a revolution in memory and storage technology, offering a unique combination of affordable large capacity and data persistence. It enables you to overhaul your storage strategy, keeping more hot data closer to the processor, and achieving greater cost/performance benefits (and so lower TCO) from your storage. Designed for data-intensive workloads, it delivers breakthrough restart times for in-memory databases and reduced wait times when fetching large data sets from system storage. For real-time, in-memory analytics workloads, high-capacity, low-latency, affordable cost and persistence are key (see figure 4).

**Customer Success: Evonik**

Evonik strives to be a world-wide leader in producing sophisticated specialty chemicals for a variety of customer needs including 3D printing, tires, and sustainable farming. It relies heavily on real-time analytics and reporting. To support real-time insights, it needed to create greater in-memory database capacity without sacrificing time or cost.

The organization wanted to reach the next level in total cost of ownership (TCO) efficiency. To reach this goal, Evonik worked with Accenture and Intel to implement servers using 2nd Generation Intel Xeon Scalable processors with Intel Optane DC persistent memory. Previously, Evoniks’ sole option for increasing memory capacity was limited to investing in larger servers. With Intel Optane DC persistent memory, it can now invest in persistent memory modules. Doing so gives Evonik the flexibility to integrate data sets into its SAP HANA platform more efficiently.

In a recent proof of concept with Accenture, Intel and SAP, Evonik found that with Intel Optane DC persistent memory, the company could save time during data table reloads after the server was restarted. Faster data loading at startup allows for shorter maintenance for SAP HANA patching or configuration changes. In a stable SAP HANA environment with a large memory footprint that supports both Intel Optane DC persistent memory and DRAM, Evonik can achieve a lower TCO. With less server downtime for its SAP HANA systems, there is also more time for productivity.
Finally, data warehouses or data lakes based on platforms like Cloudera® are common, and most organizations tend to hold a mix of data lakes and the data islands described above. Data lakes can be built on physical servers, or in the cloud or on VMs and containers, but should only be created where they support a specified business need. We are seeing an increasing trend towards the disaggregation of compute and storage. This model works well for real-time, edge-based analytics workloads, when supported by a clear storage strategy. Set criteria for tiering your data depending on its urgency and usefulness, and ensure you choose the most cost-efficient memory or storage technology for each tier.

Here again, QLC NAND and SSD technologies like Intel Optane DC SSDs could be a good fit to support high data volumes and fast caching, at a relatively low cost.

Intel® Optane™ DC SSDs remove performance bottlenecks by delivering high throughput at low queue depths, ideal for caching of temporary data in big data workloads. They deliver improved read latency performance for faster and more consistent time to data analytics insight.

AVOID COMMON MISTAKES WHEN SETTING YOUR STORAGE STRATEGY

- Choose your tools based on your data’s characteristics and needs (source, type, latency etc), and work with data gravity as much as possible – bringing compute to the data rather than always moving all the data to a centralized compute resource.
- Prepare your data carefully and think through which elements will need to come together to deliver the insights you need. They may sit across on-prem, private and public cloud so ensure these environments can work together.
- Be as realistic as you can about data volume – a year (or few) out as well as today. Plan to build in the storage, compute and network capabilities you’ll need before you hit a bottleneck.
- Keep your data as close to your compute engine as possible to avoid starving the CPU and creating cost inefficiencies for your core analytics applications.
- Be clear on where each type of data needs to sit on the storage/memory continuum (see figure 4). For example, data feeding real-time applications running on SAP HANA would need to sit in memory to deliver the low latency needed, so be sure you’re able to support and fund potentially high data volumes in-memory. Meanwhile data that’s processed in large volumes across multiple clusters (for example, using Hadoop®) would sit on cost-optimized SSD or hard disk, a lower storage tier. Consider what your future use of metadata and cataloguing might be that could stress your NAND SSD device.

Figure 4: The storage/memory continuum
When preparing for big data analytics and AI workloads, you’re likely to use the ELT model outlined above, so again compute, storage and network capabilities are important considerations. Determine not only what you need today, but how you’ll need to scale your capabilities over time. The large volumes of real-time data that are hallmarks of big data analytics-type workloads generally need a scale out rather than scale up infrastructure to keep them functioning at the required performance levels as the data grows.

Most organizations using traditional business intelligence (BI) practices today, will evolve towards more sophisticated AI techniques like machine and deep learning over time, so ensure the tools you choose will see you through.

End-to-end analytics platforms that cover all four stages of the data pipeline, like SAP*, Oracle* or Microsoft*, typically run mission-critical enterprise workloads and allow for easier integration of your business data. Intel works closely with these and other specialist solution providers and key ecosystem players to help those organizations keen to hit the ground running and achieve optimal performance.

You can also make use of AI-optimized hardware to boost your processing capabilities. The new 2nd Generation Intel Xeon Scalable processor can provide the right foundation for your analytics and AI-ready infrastructure. This proven, world-class, and highly integrated platform delivers highly scalable performance for a wide variety of analytics, AI, and other data center workloads. It also offers low TCO for batch deep learning inference and classic machine learning workloads.

Large-scale data processing, using tools like Spark*, Flink* or MapReduce*, involves reading data from multiple sources and then normalizing or aggregating that data into a consistent, readable format for the target application(s). Often this involves more data than can fit on a single machine, so you can use the frameworks to manage distributed compute clusters and provide software tools to help.

Meanwhile, the 2nd generation Intel Xeon Scalable processor and Intel Optane DC persistent memory combine to deliver accelerated insights on in-memory and other analytics workloads, with up to 8x performance improvement on queries². For example, they offer up to 2.4x faster time to insights when running SAP HANA³ compared to a four-year-old system. Meanwhile, 25Gb Intel® Ethernet 700 Series connectivity, combined with the latest generation processor technology delivers business insights up to 2.5x faster compared to 1Gb Ethernet products⁴.

Let’s take as an example a large media company focused on improving total cost of ownership (TCO) for the company’s massive content delivery infrastructure. It needs to improve storage performance and lower the cost of memory while increasing its overall memory footprint. It selects 2nd
generation Intel Xeon Scalable processors and Intel Optane DC persistent memory for its Hadoop platform. As a result, it can reduce the number of servers, improve response time and lower TCO, making it simpler for both the site of data center operation as well as the production engineers to run and operate the application.

Further accelerations can be achieved with additional technologies such as Intel® Field-Programmable Gate Arrays (Intel® FPGAs), which enable customized acceleration for a range of compute-intensive applications; and Intel® Quick Assist Technology (Intel® QAT), which provides a software-enabled foundation for security, authentication, and compression, significantly the performance and efficiency of standard platform solutions in a software-defined infrastructure.

Make use of software optimizations and frameworks to help you get your new deep learning workloads up and running quickly. Common frameworks like TensorFlow*, Caffe*, PyTorch*, BigDL* and MXNet*, and platforms like Analytics Zoo*, are designed to help you deploy machine learning or deep learning at speed, so explore which ones are most suited based on your data needs. As always, focus on the use cases that will deliver the strongest, most economical value for your organization: there's no need to go straight for deep learning when existing analytics techniques have yet been used to mine your data's fullest potential.

Intel has worked closely with the software ecosystem to develop libraries, toolkits and software optimizations that help your developers boost app performance and cut development time, for little or no cost. For example:

- The Intel® Data Plane Development Kit (Intel® DPDK) is a set of libraries and drivers that accelerate packet processing, available for free, which can be used with any x86 platform.
The new 2nd Generation Intel® Xeon® Scalable processors with Intel® Deep Learning Boost accelerates AI inference up to 30x—helping you get the most out of your investment⁵.

- Intel® Distribution of OpenVINO™ toolkit enables developers to quickly build applications and solutions that emulate human vision.
- The Intel® Math Kernal Library (Intel® MKL) optimizes code with minimal effort for future generations of Intel® processors.
- The Intel® Data Analytics Acceleration Library (Intel® DAAL) helps software developers reduce the time it takes to develop high-performance applications, boosting machine learning and data analytics performance in these applications.

AVOID COMMON MISTAKES IN SETTING PROCESSING STRATEGY

- Make sure you’re aware of all the tools and ingredients available to you — including software optimizations and frameworks. Choose those that support the most valuable use case for your business.
- IT organizations used to using CPUs may feel pressured by developers or users to invest in GPUs to support big data analytics and AI. However, while GPUs work well when running a large number of computations on small amounts of data, they are not suited to database operations that require a lot of data to be moved back and forth. This type of workload — common in analytics — may be best served by a CPU-based environment like the one you’re already running.
- Remember memory. Many real-time analytics use cases (such as fraud monitoring) rely on in-memory databases that can quickly run out of capacity if sufficient memory is not available in the system.
Stage 4: Act

At last we reach the analysis stage of the data pipeline, where we pull actionable insights from the data we have now integrated, stored, cleaned and processed. This stage involves in-depth exploration and visualization to better understand the results.

It is at this stage that experts like data scientists and business analysts must play a role in helping turn these insights into business value. The techniques they deploy here will vary depending on the project. They will perform a range of aggregation and analysis tasks across multiple datasets, which may vary widely in type and structure. Possible techniques include statistical and machine learning methods, like linear/non-linear modeling, clustering techniques, dimensionality-reduction techniques, graph-based methods, and neural networks.

During the insights phase, data visualizations, dashboards and reports can help business users understand the results. The software you use for this should be flexible enough to evolve and develop over time in line with changing business needs and new analytics and AI use cases. To make this easier, popular cloud service providers integrate with a number of reporting and dashboarding tools.

It is important not to set unachievable goals. Begin with your existing infrastructure and business intelligence (BI) tools, then evolve as your capabilities and resources expand. For example, an online retailer may begin by using structured data in a relational database to track customer transactions. The next step may be to complement customer information with unstructured data which would enable the organization to build a fuller picture of each customer's experience. For instance a customer looking to place an order online who gets an error message, may then phone customer support and wait 15 minutes for an answer before giving up. They may then post a comment on Twitter*. In this instance, the initial issue online is caused by a web server error, while the support call goes into a relational customer relationship management (CRM) system. By bringing these together with the social media post, the organization can gain a more complete, nuanced understanding of the customer and their pain points. This type of user behavior analytics is a great way of extracting more insight from your data over time, and laying the foundation for evolving smoothly towards more complex AI techniques like machine or deep learning.

Avoid the mistakes of setting an analytics strategy

- Ensure your analytics use case is clearly associated with an existing business process/application and will deliver tangible business value. All too often, a data scientist creates an amazing proof of concept of an analytics model on their laptop, but without clear alignment to a business need, the model will never go mainstream within the organization.

- Pick the right problem. Choose a starting project that will deliver value quickly, without getting into too much technical debt. Using existing infrastructure and/or traditional BI tools to run your first analytics projects will help prove ROI before you invest in new technology.

- Don’t forget to optimize your analytics models over time. In addition to business insights, you can gather information that will help you enhance the velocity or volume of data ingestion, the use of different storage mediums to speed analysis, and enhancements to the processing pipeline.
PRODUCTIONIZING: MAKING DATA ANALYTICS PART OF THE COMPANY’S DNA

The key to productionizing analytics is getting the business to adopt it, and this depends on both business users’ willingness to adopt it, and IT’s ability to scale.

In this way, the end of this journey comes back to the beginning – picking the right use case and team to deliver business value. If business users do not adopt and incorporate the analytics into their daily routine to make insights-driven decisions, the analytics project is a failure. The metrics identified in your foundational efforts (such as improved efficiency, reduced waste, increased revenues) should help measure the impact of the analytics solution on your business goals, and promote adoption.

Make adoption as easy as you can for users by providing visualizations and insights that are easy to understand and apply. Visualizations should give the right information at the right time in a consumable format. Ideally, this should be embedded in the workflow to help encourage business adoption. Also, remember that the data must be accurate – if users cannot trust the data (which typically happens if short-cuts or technical debt are used to put half-baked projects into production), it won’t be adopted. Ultimately, this is why it is so critical that IT is involved in selecting the right team for analytics initiatives.

On the IT side, ensure you are equipped to deliver your newly proven analytics use cases to a larger, more dispersed user base, even as your data continues to grow. A multi-cloud strategy that takes advantage of on-premises, private and public cloud resources is important here. Intel works closely with the ecosystem to provide tools and solutions that help with this, such as a newly announced Intel® Select Solutions with Google Anthos*, which enables you to make use of existing on-premises investments while providing a clear path to the cloud. Meanwhile, containers and tools like Kubernetes* and Docker* now make it easy to shift and integrate applications, enabling your developers to use VMs and containers to build and scale more innovative apps, more quickly.

Review your existing infrastructure to identify where you’re able to meet your objectives with what you have, and where it makes most sense to invest in new technologies. Intel works closely with its ecosystem to deliver a range of Intel Select Solutions, which offer combinations of hardware, software, networking and storage that are optimized for data-intensive analytics and AI workloads. These include: Intel® Select Solutions for SQL Server (Microsoft and Linux* versions) and new additions Intel® Select Solutions for SAP HANA*, and IBM Cloud Private for Data* (ICPD*). We have spent years working together to optimize and certify key workloads and applications on Intel® architecture. We understand that whether you’re running them on-premises, in the cloud or (increasingly) moving them between the two, eliminating siloes and creating one constant and cohesive environment across all the parts of your IT infrastructure is essential.

It’s important not to be tempted to aim straight for the latest ‘shiny object’ (deep learning being a current example). By doing this, you may be leaving money on the table that could be picked up using a ‘crawl, walk, run’ approach. If external expertise is needed, consult a system integrator. Additionally many of the larger cloud service providers now offer tools to help you achieve results more quickly, without going for the most exotic solution (for example, Google’s* recent acquisition of Looker*, or Salesforce.com’s* acquisition of Tableau*).

Keep on top of maintenance and governance of your data over time as well. Monitor your entire pipeline and data usage to track trends, bottlenecks and opportunities. Apply upgrades and tweaks regularly to ensure the whole piece continues to function optimally and TCO is kept to a minimum.
While addressing these challenges may seem like a large mountain to climb, now is the time to act. A number of key systems, like SAP Enterprise Resource Planning*, Microsoft Windows 2008* and Server 2008* are due to stop offering support in the coming years. This means organizations that use these platforms will need to migrate to newer options, which have ongoing support, but also the capacity and performance needed to handle the ongoing growth of data volumes.

In getting ready for the data era, you’ll need to address some significant challenges, but by taking the approach outlined in this paper you can turn them into big opportunities. The key is to consider your data pipeline and analytics and AI needs holistically, and in common with your overall business requirements.

**CONCLUSION**

**FURTHER READING**

- Evonik Case Study: Defining the future of in-memory database computing
- Workbook: Analytics/AI Organizational Evaluation
- White Paper: Unleashing the Power of In Memory Computing: Intel Optane DC persistent memory on SAP HANA
- Technology Brief: Intel Optane Technology: Memory or Storage?
- Intel® Select Solutions for Analytics
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operating systems, and workloads. Any change to any of those factors may result in the performance tests to vary. You should consult other performance information and benchmark results for the specific computer configurations you are considering for other information that may affect performance.

1. Intel’s performance tests, which include results stated as improvements in performance or performance benefits, are measured relative to a specific baseline of performance and may vary among other Intel microprocessors or future microprocessors. Performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operating systems, and workloads. Any change to any of those factors may result in the performance tests to vary. You should consult other performance information and benchmark results for the specific computer configurations you are considering for other information that may affect performance.

2. 2.4X better runtime performance: results are based on testing by Intel [as of March 12, 2019]. No system can be absolutely secure. Baseline: three (1 master + 2 slave) SAP HANA 2 scale-out configuration. Per Node: 4x Intel® Xeon® processor E7-8880 v3 (2.3 GHz, 192 cores), 4x CPU sockets; 4x Memory: 64 x 2brig DIMM, RAM model: DDR4-2133 MHz, Storage: GFS, approximately 21.8TB of formatted local storage per node, SAN storage for backup configuration: redundant 10G network for storage and access, redundant 10G network for node-to-node; OS: SUSE SLES 12 SP2, SAP HANA 2.0, GFS, 4.4.3.10. Average time of 50 individual test queries executed 30-50 times each, for a total of approximately 25,000 steps: 2.81 seconds. New configuration, one master node SAP HANA 2 scale-out configuration: CPU: 4x Intel® Xeon® Platinum 8180 processor (2.2 GHz, 165 W, 24 cores), CPU sockets: 4x Memory: 40000011c, RAM capacity: 24 x 64GB DIMM, RAM model: DDR4-2133 MHz, Storage: GFS, approximately 21.8TB of formatted local storage per node, SAN storage for backup configuration: redundant 10G network for storage and access, redundant 10G network for node-to-node; OS: SUSE SLES 12 SP2, SAP HANA 2.0, GFS, 4.4.3.10. Average time of 50 individual test queries executed 30-50 times each, for a total of approximately 25,000 steps: 11.33 seconds.

3. Testing based on the 2nd Generation Intel® Xeon® Platinum 8260 processor and upgrading from a 1Gbps Intel® Ethernet Network Adapter XXV710 and from Serial ATA (ATA) drives to the NVMe Express(TM) based PCIe® SSD DC P4600. Performance results by Headstart Strategic Communications are based on testing as of February 12, 2019. The comparative analysis in this document was done by Headstart Strategic Communications and commissioned by Intel. Detailed configuration: VM Host Server Intel® Xeon® Platinum 8160 processor Intel® Xeon® Platinum 8160 processor (CPUID 0x68, memory revision 0x00004), and Intel Xeon® Platinum 8160 processor (CPUID 0x68, memory revision 0x00004), Intel® Server Board S200W88W, Intel® Xeon® Processor E5-2680 v3 (14 cores, 20MB cache, 2.50GHz, 85W), Intel® Xeon® Processor E5-2680 v3 (14 cores, 20MB cache, 2.50GHz, 85W), and Intel® Xeon® Processor E5-2680 v3 (14 cores, 20MB cache, 2.50GHz, 85W). Intel® Optane DC persistent memory: 24 x 126GB PMM; storage: XFS, 21TB; network: redundant 10G network for storage and access, redundant 10G network for node-to-node; OS: SUSE SLES 12 SP2, SAP HANA 2.0, GFS, 4.4.3.10. Average time of 50 individual test queries executed 30-50 times each, for a total of approximately 25,000 steps: 2.81 seconds. New configuration, one master node SAP HANA 2 scale-out configuration: CPU: 4x Intel® Xeon® Platinum 8180 processor (2.2 GHz, 165 W, 24 cores), CPU sockets: 4x Memory: 40000011c, RAM capacity: 24 x 64GB DIMM, RAM model: DDR4-2133 MHz, Storage: GFS, approximately 21.8TB of formatted local storage per node, SAN storage for backup configuration: redundant 10G network for storage and access, redundant 10G network for node-to-node; OS: SUSE SLES 12 SP2, SAP HANA 2.0, GFS, 4.4.3.10. Average time of 50 individual test queries executed 30-50 times each, for a total of approximately 25,000 steps: 11.33 seconds.

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice Revision #20110804

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data are accurate.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

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 intel.com

Intel, Xeon, Optane, OpenVINO, and the Intel logo, are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

© Intel Corporation 0719/1/P/VCAT/PDF

340894-001EN