SAP HANA's Defining Capabilities

The Best-Run Businesses Run SAP™

Table of Contents

Overview

SAP HANA with its new vision and approach, as with many things novel and paradigm-shifting, can be met with resistance and disbelief. Change is not always easy and many who are comfortable with existing approaches try to compare SAP HANA with, and to see it in the light of, those older paradigms. This, however, cannot successfully be done given that SAP HANA represents a completely new way to address traditional computing issues. There is a SAP HANA difference and this paper will focus on some of the unique capabilities of SAP HANA's revolutionary approach — especially the technological underpinnings — and how these translates into a very different way of assessing value for those who deploy it. Some of these unique SAP HANA capabilities that will be explored here include:

- SAP HANA's **Data Locality Aware Processing.** SAP HANA's CPU-aware optimizer leverages optimizations of the data's locality to the CPUs and their caches to reduce latency of data compute. SAP HANA ensures that data resides in the DRAM and CPU caches to avoid cache hit misses, and applies knowledge of all the bandwidth and latencies between every cache, every CPU, and every node in the execution plans.
- SAP HANA can run real-time **OLTP and OLAP on a single copy of the data.** There is no need for unique OLTP and OLAP copies. Data is stored only once for both transactional and analytical applications.
- Multiple Parallel Processing **(MPP) on a shared nothing architecture along with its support for Single Instruction, Multiple Data (SIMD).**
- Even though the SAP HANA true in-memory platform is a paradigm-shifting new approach, nothing has changed with regards to ACID compliance or Reliability, Full High Availability, Disaster Recovery and Supportability. For all its cutting-edge innovations, it is still ACID compliant and it supports Disaster Recovery (DR) via synchronous or asynchronous native replication.
- Unique **on-the-fly schema extension capability** allows for flexible business model changes.
- **Dynamic Data Tiering** optimizing the balance between data processing and data storage. Data identified as frequently used is kept in memory. Any data not recently accessed is purged from memory but persisted on-disk without write-back.
- It is **more than just a database.** SAP HANA converges platform, database, data processing capabilities, handles spatial and textual data analysis, and provides libraries for predictive, planning and business analytics. All in a single platform with a built-in application and web server.
- SAP HANA provides the most comprehensive data provisioning for any data from any source.
- It is **open and agnostic** meaning any application and any data. SAP HANA runs on commodityx86 based hardware from SAP partners. It supports a wide variety of languages, certified 3rd party tools, and custom apps created in almost any language. It is data agnostic with support for structured, unstructured (i.e. text), spatial, document data, and sparse data.
- SAP HANA has **extreme linear scalability** scaling up and scaling out. Adding more nodes to handle more data, keeping response time relatively unchanged.

Let's examine these unique and game-changing capabilities in more detail.

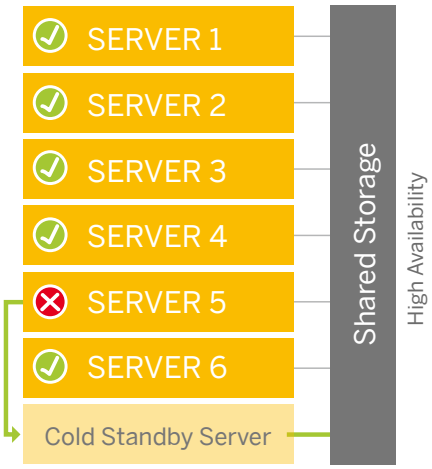
The SAP HANA Difference

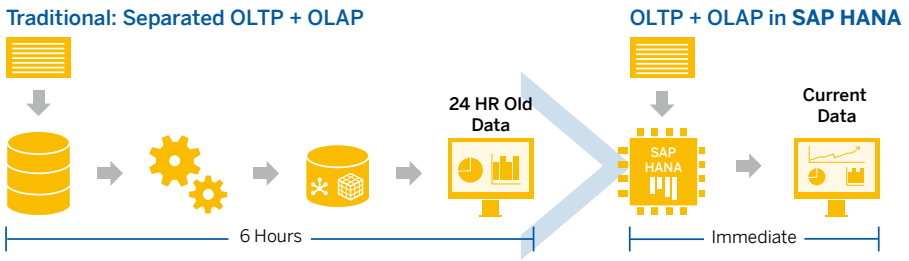
SAP HANA is more than just a database, but at its heart there is columnar database with a data processing engine at its core. And one that is constantly evolving. Two years ago, when it first came out, SAP HANA was capable of scanning data at 2.0 billion integers per second per core, already an impressive rate. Today, this has increased to 3.2 billion integers per second, per core. This continued drive for improvement is important because, all data in SAP HANA gets represented as integers. This means all data is capable of being scanned at this incredible speed. Scanning, while it is a good speed test, does not indicate actual processing speed. SAP HANA is also capable of handling 12.5 million aggregations per second per core. A standard SAP HANA node (i.e. single server) has 40 cores. Meaning that a single SAP HANA node can scan data at 128 billion integers per second and aggregate data at 158 million integers per second. Where this gets really interesting is realizing that SAP HANA can ingest data at 1.5 million records (i.e. rows of data) per second per node, all the while maintaining and supporting all the ACID properties.

- Atomicity: Requires that each transaction is "all or nothing". If one part of the transaction fails, the entire transaction fails, and the database state is left unchanged.
- Consistency: Ensures that any transaction will bring the database from one valid state to another. Any data written to the database must be valid according to all defined rules, including but not limited to, constraints, cascades, triggers, and any combination thereof.
- Isolation: Ensures that the concurrent execution of transactions results in a system state that would be obtained if transactions were executed serially.
- Durability: Means that once a transaction has been committed, it will remain so, even in the event of power loss, crashes, or errors.

So while it can be said that SAP HANA is a database capable of running any application that a relational database can run, that is just the beginning. Before you read any further, please look at the 'Durability' point in the ACID list just above and recognize that SAP HANA is ACID compliant and does not lose data in event of a power loss (or any other type of failure). Just like any other ACID compliant database, SAP HANA writes any changes to its data to non-volatile memory before acknowledging success. SAP has 'Durability' and it is fully recoverable.

SAP HANA is more than just a database. The danger of viewing SAP HANA simply as a database is that it associates SAP HANA with the complexity of most databases. Most databases have the need for tuning with indexes, specialty purchased and task configured hardware, and creation of aggregate tables to maintain the speed of transactional systems or to "pre-answer" queries in analytic ones. But SAP HANA has none of these drawbacks. The speed of SAP HANA, while game-changing and beneficial (cost of FTEs staring at hour glasses adds up), allows it to run without indexes, aggregates, or specially configured hardware and software. You don't need to create indexes to make SAP HANA achieve the astounding read speeds noted earlier in

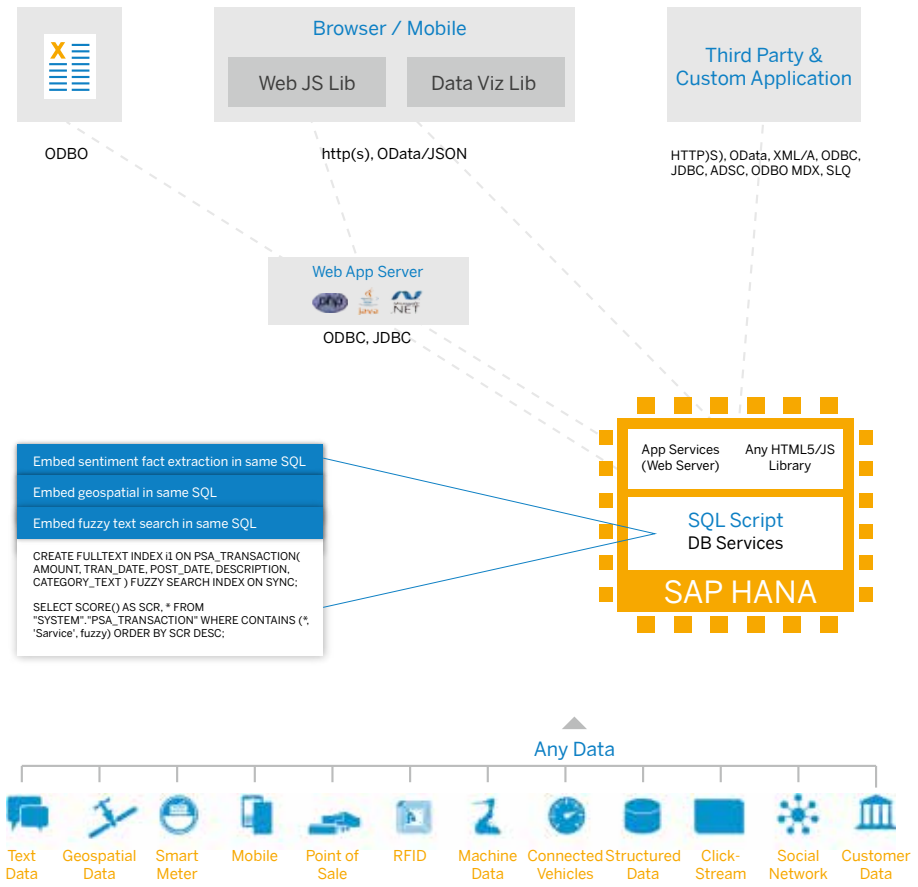




The SAP HANA Difference

the paper. You don't need to create materialized aggregates to make the system achieve acceptable performance. You don't need to configure your hardware and database based on whether the application will be used for OLTP versus OLAP (a.k.a. analytics). Just load data into SAP HANA (or create virtualized tables to access data in remote source systems locally) and it will run faster than any other database – period. This also means that it costs less to run, but again more on that later.

One of the benefits of SAP HANA's speed is that it is schema agnostic. What this means is that your OLTP schema doesn't need denormalization in order to perform. This reduces development time and makes the schema self-describing and thus supportive of agile methodologies. Schema agnostic in SAP HANA also means that you can do full analytics on this same schema. Let this sink in. You no longer need to ETL data into an operational data store in order to off-load processing from the OLTP system. That is half the hardware costs and requires no administrators to monitor and maintain a second system. Additionally, analytics data is now 100% real-time – no batch. Need to know what orders are open — just ask. Note that other vendors entering the market are making similar sounding claims, but with dramatically fewer benefits. They are putting very fast columnar tables alongside their OLTP row tables. This is good. A second server for operational reporting is no longer needed. That said, the rest of the benefits SAP HANA provides by being schema agnostic are lost in these competitive systems. The OLTP data still needs to be migrated to the read-optimized columnar tables and thus losing the real-time benefit. Additional storage is needed as the data is duplicated. Administrative costs are higher as you need to monitor and administer the change data capture process on the row tables and the delta data activation process on the columnar store. How does SAP HANA avoid all this and provide superior benefits? Simple. SAP HANA's table structure has the OLTP speed of row tables and the read speed of columnar tables. There is no cut-over needed, no row table to



SAP HANA's Defining Capabilities

Easily migrate your applications
(e.g.L: Java, PHP, .NET) in almost any language, PHP, Ruby, Java, C,...the list goes on:

- Support for ANSI SQL, ODBC, JDBC, Odata/JSON, and certified 3rd party tools.
- Support more standards: JSON and XMLA over HTTP so it is truly multi-dimensional platform.

Build new web applications with any open source
HTML5/JS libraries, Server Side Java Script.

Open Cloud Partner Program
Allows you to select the best SAP HANA cloud option from partners.

Support advanced text analytics:
Analyze text in all columns of table and text inside binary files with advanced text analytic capabilities such as: automatically detecting 31 languages, fuzzy, linguistics, synonymous search, using SQL.

Analyze streaming data
From integrated ESP in combination with data in SAP HANA.

Process geospatial data

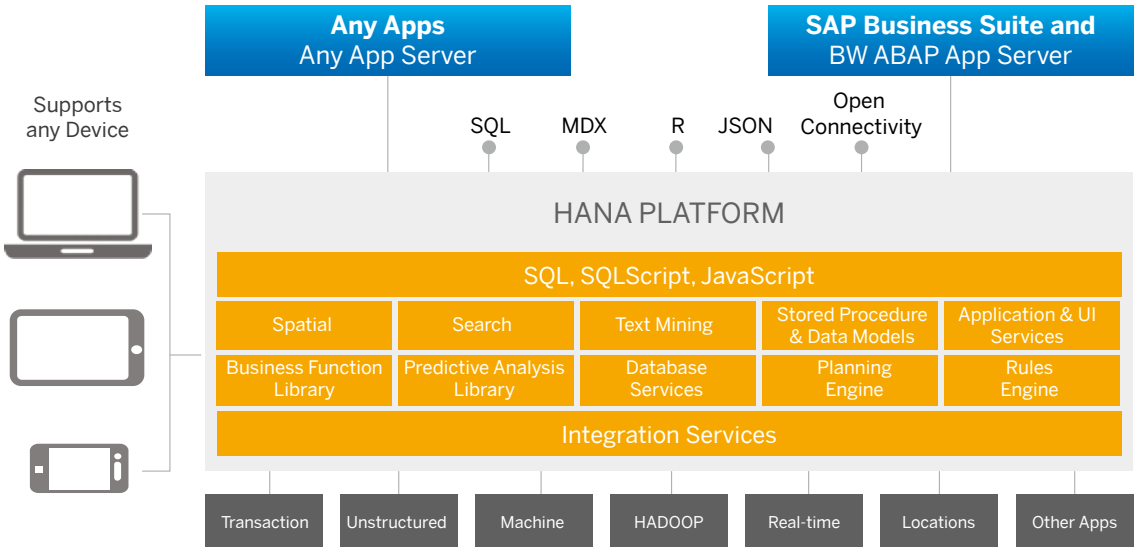
The SAP HANA Difference

columnar table conversion needed, no batch building for columnar compression, no nothing. The same table supports OLTP and OLAP Analytics – at the same time – with full ACID properties. This means real real-time and real-value.

Another advantage SAP HANA has over traditional databases is openness. SAP HANA is open and agnostic. It supports ANSI compliant SQL, ODBC and JDBC protocols, and a wide range of certified 3rd party query tools, ETL tools, back-up and recovery tools, monitoring tools, and even applications. Yes, even ones that compete directly with SAP applications. SAP HANA supports custom applications built on SAP HANA in any language including PHP, Ruby, Python, Java, C, and of course, ABAP. And SAP HANA also supports MDX over ODBO and XMLA over HTTP if you want to access SAP HANA as a multi-dimensional database. SAP HANA is data agnostic as well, not only supporting tables, but unstructured data (i.e. text), spatial data, document data, and sparse data as well. And SAP is getting ready to release support for property graph data in SAP HANA as well. With unstructured data, SAP HANA supports ingestion of a huge number of document types including all the usual suspects, language identification for 31 languages, entity extraction for three languages, and sentiment analysis for the top five spoken languages. For spatial data, SAP HANA is not only screaming fast but also ships with a geo-coding service, polygons for all major political and geographic areas, and an industry compliant integration layer supporting ESRI, Navtec, and Google. As for document data, SAP HANA supports complete transactions using JSON or XML. For sparse data with varying attributes, SAP HANA supports flexible tables in which columns can be dynamically added on Insert or Update. All vitally important non-structured data types explode.

There are several new specialized databases built specifically for these data types that have allowed for applications to address the 80% of all data that isn't structured (i.e. not relational). These purpose built databases have provided, huge innovation leaps. But using them is still a separate database approach with all its inherent issues: The need for separate servers, data to be copied between the systems, and applications having to merge the data coming from the different systems via in-app code. All a lot of work and effort. This may be fine if you are a specialized web start-up, but definitely not if you are an enterprise. Additionally, these complex infrastructures reduce agility. By being in the single SAP HANA data platform, developers can easily take advantage of data of any kind to build next generation applications. Data of heterogeneous data types can be easily combined, processed, and leveraged.

SAP HANA also brings the analytics code closer to the data. In and of itself, this isn't new or revolutionary. Stored procedures have been around forever. The unique aspect is the offering and provision



The SAP HANA Difference

of pre-built libraries and the speed with which SAP HANA can run the analytics. There is a Predictive Analysis Library (PAL) which supports 27 data preparation and data mining functions including clustering and classification algorithms, regression methods, decision trees, and forecasting methods. These are all natively compiled and parallelized as part of the query execution plan determined by the SAP HANA optimizer providing blinding speed. There is also a Business Function Library (BFL) that contains 55 data intensive functions that would slow down traditional three tier applications given their need for data movement from the database tier to the app tier. Things like currency conversion, depreciation, and moving median calculations. And new libraries are being added to SAP HANA for data quality and data transformations. There is also a high speed integration to "R" that lets you utilize the over 3500 algorithms that have been built on this open source analytics platform to leverage the real-time capability.

Individually these innovations are great, but what is really amazing is that that you can use them together. SAP HANA makes this not only possible, but even easy. Want to show a billion pieces of data on a map showing visualizations based on clustering analysis of that data? How about a financial system that takes every transaction for the quarter up until this last second, does a synthetic close on the books, and then forecasts where the financials will be at quarter end?

So SAP HANA is a fully ACID compliant database that supports real-time OLTP and OLAP (analytics) without duplicate copies of the data. It supports every major data type along with text analytics and geo-spatial capabilities. It supports all major industry standard query languages, and comes with highly optimized industry functions and algorithms to build class best performance (think 1000s of times faster). And it does this without indexes or aggregates. When these are combined in totality, you have a system that uses less hardware and requires less maintenance and saves you money while providing game-changing capabilities and agility.

Architecture

SAP HANA's speed is often referred to as being "in-memory", where all data can be placed in the computer's RAM and thus avoid slow disk reads. This may be true, but overshadows other huge innovations developed by the HANA team in partnership with Intel. These innovations deliver orders of magnitude faster performance over just putting data a big cache.

SAP HANA's version of in-memory works by:

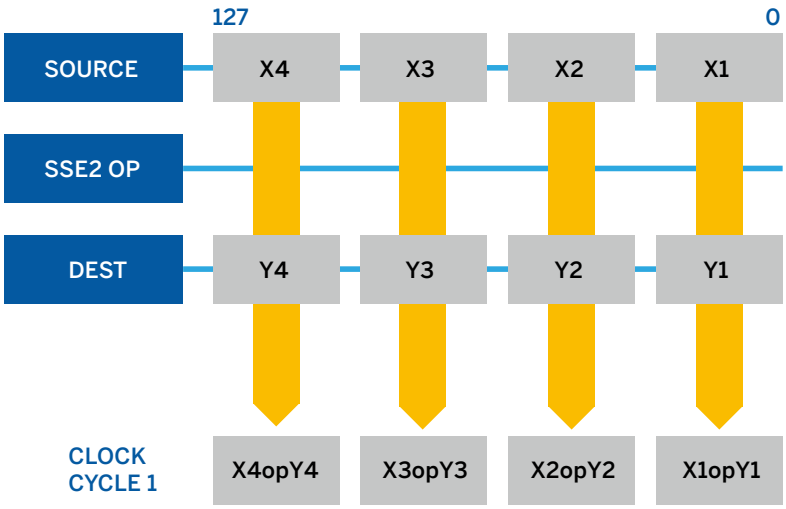
1. Having a SAP HANA optimizer that leverages knowledge of the data's locality to the CPUs and their caches. It leverages this information along with knowledge of all the bandwidth and latencies between every cache, every core, every CPU, and every node to formulate the most performant execution plans
2. Aligning all data in-memory with the processor's cache-line size (the amount of data always fetched by the processor's memory controller). This allows data to be pipelined from RAM, through the caches, and into a core's registers.
3. Aligning data in-memory in vectors (columns) allowing data to be compressed highly efficiently by data type. Therefore SAP HANA is able to pipeline much larger amounts of data to the core's registers.
4. SAP HANA acts on compressed data whenever possible including scans, aggregates, and joins.

When processing data, SAP HANA leverages Intel's SSE vector processing technology (i.e. SIMD instruction set) to process multiple pieces of data in a single instruction cycle. Since SAP HANA uses variable length integer encoding, this means that for lower cardinality columns, it can process 10s of integers per register per instruction cycle.

Now that we've established where SAP HANA's speed comes from, we can address perceived concerns around out-of-memory conditions, scalability, and data center readiness. SAP HANA is resilient. When available memory is filling up, SAP HANA purges any column partition, column, or table that hasn't been accessed recently from memory. It is not lost. This is efficient as the data is already persisted on-disk and no write-back is needed. BLOBs really don't need to be in-memory once their entities have been extracted, so these can be set to only reside on disk saving memory. In most instances SAP recommends leaving all data in memory, but you don't have to and be less frequently accessed data can be stored in systems such as Hadoop and SAP Sybase IQ. Further SAP HANA also doesn't need that much memory to process very large amounts of data. Most DBAs quote their databases in total size, including indexes and aggregates. So a 1TB database typically only has in approximately 500 GB of raw data. SAP HANA achieves compression of OLTP data in the 3-6X range and data warehousing data in the 10-20X range. To be conservative, let's assume 5X compression. This makes the 500GB of raw data 100 GBs. Now, SAP HANA doesn't only store data in-memory, it also processes all its interim datasets in-memory, so we recommend a SAP HANA instance memory of 2X the compressed raw data, or 200 GB in our example. This means that you can get all the benefits SAP HANA offers for a 1TB existing database in a 200GB SAP HANA node. Said another way, a standard 1TB SAP HANA node is capable of handling a 5TB database on migration. SAP HANA is also fully elastically scalable allowing you to add nodes to accommodate more memory and data. There are certified configurations up to 56 nodes and 56 TB of memory and SAP has built a 100 node 100 TB system on which a 1 PB benchmark (20x compression) was run. Conceive of needing even more than 56 nodes? Just ask and SAP will certify more.

What is truly incredible is the linearity SAP HANA exhibited during scaling tests

Single Node SIMD Operation



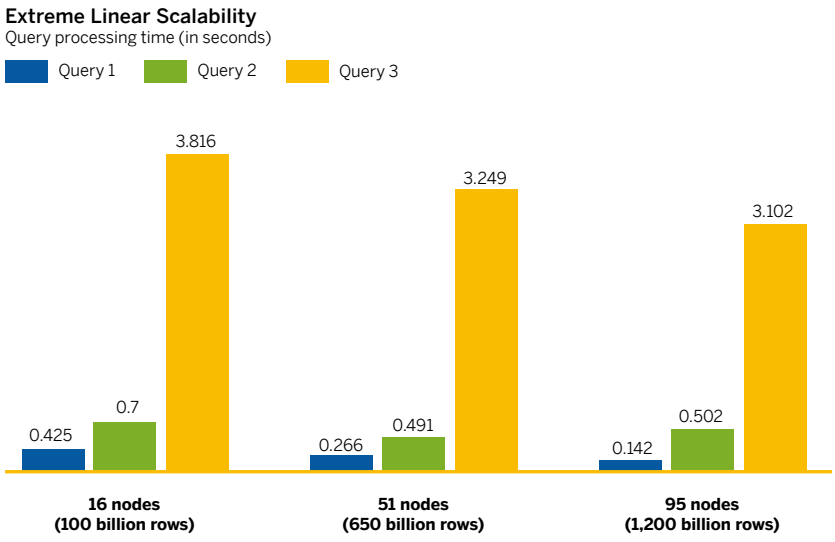
Architecture

on the 100 node system noted above.

As you can see from the graph above, SAP HANA's response time was essentially flat as data and nodes were doubled. Simply stated, scaling SAP HANA just takes math. Want to double the speed of your queries? Double the nodes? Want to double your users and maintain response times on the same data size? This is because SAP HANA is able to fully parallelize any query and fully saturate its CPU cores. This means that SAP HANA is fully utilizing all available resources while wasting none. It means that SAP HANA is always going to solve your query as fast as is possible. If you have large concurrency on SAP HANA, its execution plans will use fewer cores and memory to allow all users to be serviced. And a planned innovation will make SAP HANA even faster. As noted by Hasso Plattner at this past year's Sapphire conference in Orlando, FL, SAP has developed a new object aware cache that is capable of handling delta changes. This means that similar queries should get millisecond response times even under load once this innovation is released.

SAP HANA offers reliability, full high availability, disaster recovery (DR) and supportability. So from a data center readiness perspective, SAP HANA is complete. SAP HANA fully supports fail-over nodes so that if a node fails for any reason a backup node will kick in. Since SAP HANA is in-memory, performance will be slowed to disk-database speeds until all data is moved from disk to memory. Note that the fail-over node is available immediately and data will load as queries require it. SAP HANA also supports DR via either synchronous or asynchronous native replication. This allows secondary SAP HANA systems to be located continents away providing protection from any natural or political disasters. There are many options for DR that can allow for near zero down time upgrades or use of the secondary system for use as test and dev. Our partners are also offering log shipping as another DR option and efficient use of hardware.

As noted earlier, SAP HANA is open and agnostic and that goes for hardware as well. SAP HANA runs on "commodity" components assembled by our hardware partners to the minimum specifications given to them by SAP. These specifications allow very different architectures by each partner providing our customers both choice and a competitive market. We recently announced an initiative allowing customers to use their own storage systems as long as they meet SAP HANA's specifications. And SAP HANA can also be run virtualized to support testing and dev. All these innovations are designed to keep hardware costs down. Lower costs does not mean cheap or chintzy. SAP HANA hardware is not inexpensive as standard components are enterprise class and the servers do have lots of memory. But, using compression, hardware and storage can be conservatively sized. And you do get what you pay for. As an instance can handle mixed workloads of OLTP and operational reporting, one less server is required. And as SAP HANA doesn't require indexes and aggregates, data duplicates in row and columnar structures, customers report lower operating costs. So you get good value and good TCO. Take SAP's own experience in migrating its CRM system. SAP saw a 30% reduction in hardware costs due to no longer needing a separate system for operational reporting.



Sales and Distribution Reports

Query 1: Single Customer and Material for One Month

Query 2: Range of Customers and Materials for Six Months

Query 3: Year-Over-Year Trending Report for Top 100 Customers for Five Years

Summary

SAP HANA represents such a huge shift in data management, application development, and analytics that it is a paradigm shift. And like most things new and earthshattering it is often hard to believe or accept at face value, especially when looked at through the lenses of existing paradigms. SAP HANA takes the most important innovations in hardware architectures, columnar and row databases, text analytics, parallel processing, and much more and blends them together in an elegantly designed software platform and appliance. SAP HANA is convergence. SAP HANA represents a huge shift for enterprise computing. SAP HANA is the first technology in decades that will simultaneously allow landscape simplification (i.e. fewer servers) and rapid innovation. Said another way, SAP HANA is the next generation data platform.

www.sap.com/contactsap

13/08

The Best-Run Businesses Run SAP™

© 2013 SAP AG or an SAP affiliate company. All rights reserved.