# SAP® HANA™ ESSENTIALS

Jeffrey Word

Book Excerpt: Introduction to SAP Big Data Technologies. To get the complete book for free, go to http://saphanabook.com and use code 6362EC9A

The author and publisher gratefully acknowledge SAP's kind permission to use its trademarks in this publication.



ISBNs:
978-0-9856008-0-8 (ePub)
978-0-9856008-1-5 (Kindle)

# About the Cover Image

The cover image is a European "No Speed Limit" sign. If you've ever driven on the Autobahn in Germany, this sign will immediately bring a smile to your face because you can step on the accelerator and drive as fast as you want to or as fast as your car can go (which ever comes first). In terms of SAP HANA, we selected this image because SAP HANA allows your company to run at top speed with no artificial limit to how fast it can go. If you ever go visit SAP headquarters in Germany, you'll see this sign about 2 miles south of the Frankfurt airport on the A5 — and there's no speed limit on your way to visit SAP.

Chapter 6

# Introduction to SAP Big Data Technologies

*"Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world."*
*— Atul Butte, Stanford*

This chapter was written with the expert assistance of John Schitka and David Jonker, SAP Big Data Marketing.

## What is big data?

Big data is more than just data volume or size. It is about generating valuable real-time insights from raw data, no matter the size, the type, or the rate at which it is generated.

Leveraging big data, organizations are successfully uncovering new insights from all of their data, creating opportunities to transform businesses, industries, and even the quality of our lives.

Big data is not an intrinsic good in and of itself. It is useful as a stepping-stone to business value. Many organizations see the opportunities for big data and understand some of the use cases. The problem is that there are challenges in getting to that business value. In attempting to integrate big data with their existing data sources, organizations face questions and concerns such as:

- Lack of skills—Where can I find the resources to make this project a reality?
- Slow deployment—How do I speed up the implementation time, reducing the effort to implement a solution or application?
- Complex IT environments—How do I rationalize new big data technologies in an already complex IT environment?
- Integrating many data sources—What is the relationship between all of my data sources and how do I normalize that relationship?

To work with big data you need to be able to acquire it, analyze it, and act on insights derived from it.

The SAP HANA platform provides these capabilities. It delivers in-memory processing of data with tiered, petabyte scale storage and integration with SAP IQ and Hadoop, both of which will be touched on in this chapter.

This chapter explains what big data is and how you can leverage it as part of your system landscape. It describes how three groups can benefit from the big data capabilities inherent in SAP HANA:

- BI analysts. These analysts have been used to working with traditional data sources such as data warehouses and systems of record and helping organizations support a single version of the truth using SAP BusinessObjects and other BI tools.
- Analysts and data scientists using advanced analytics. Analysts and data scientists are trained to work with the variety, volume, and velocity of big data.
- Everyone through operationalized insights. Everyone in an organization benefits when insights are automatically delivered in context. Through embedded analytics, insights from big data and traditional data sources are integrated into the context of business processes and applications. In this way, the entire organization becomes more data-driven as a matter of course, capable of repeatable victories based on the latest information.

This chapter provides an introduction to SAP's big data technologies, with, as you would expect in this book, a focus on SAP HANA. Links to further resources will be provided for those interested in learning more.

## Big Data Characteristics

The journey to big data has taken us into a world where data sets are growing in size and complexity and where there are ever-increasing demands for query execution speed. These three driving factors of volume, variety, and velocity were first enumerated by Gartner's Doug Laney in 2001, and the ensuing decade has only brought them into sharper focus.

The volume of data generated by modern computing systems, sensor networks, and social media streams is ever growing. What was once considered digital exhaust, only to be collected for audit or regulatory reasons, has now become a treasure trove of information. Data storage costs have been reduced to the point where it is more cost effective to save anything that you'd ever

expect to need and sort through it later rather than spend scarce resources up front assessing its ultimate worth.

Data is being collected in a wide variety of formats, ranging from simple relationships collected by tiny machines up to complex multimedia. The modern data center houses an ever-growing range of data formats as more and more systems come online, producing transactional and analytical data in a plethora of data structures including bulky and complex voice and video formats.

The most important aspect of big data has been the need for velocity in all aspects of data management. It has long been possible to store immense volumes of data in relational systems, but query times were so slow as to be unusable for real time operations. Thus, in practice many relational databases never stored more than several terabytes of data because response times would degrade too much. Inserting even more varieties of data into that database would further degrade response times, causing IT staff to keep the database clean of all but the most important transactional information. In other words, big data is a velocity problem that is exacerbated by greater volumes and varieties of information. Immense data stores with widely varying data need to have fast performance so that complex analytical tools can turn around insights and help inform decisions quickly. http://blogs.sap.com/innovation/big-data/big-data-is-not-about-big-data-028590

### The Ultimate Goal: Faster, Data-Driven Decisionmaking

Big data is important because ultimately it can improve decisionmaking. However, time is of the essence. Figure 1 shows how decision making works and the delays inherent in it. It comes from a 2003 paper by Dr. Richard Hackathorn called *Minimizing Action Distance*.

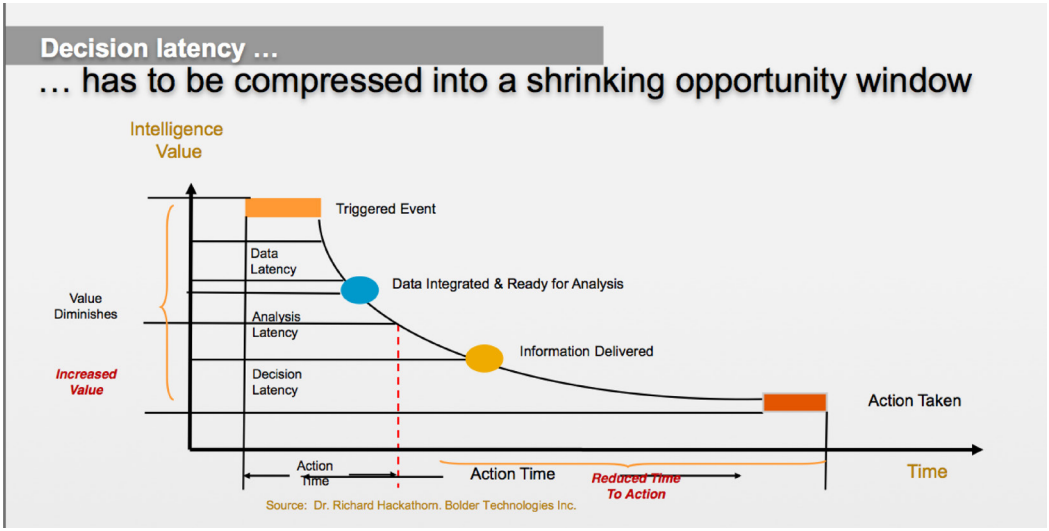



**Figure 1. Minimizing Action Distance**

Hackathorn shows that there are three types of latency in the decision making process:

- Immediately after the triggered event, there is **data latency,** where data is integrated and made ready for analysis. Sometimes this also involves several steps of preprocessing.
- After the data is prepared for analysis, there is **analysis latency**, the time involved in initiating the analysis, packaging its results, and delivering it to the appropriate person.
- After the information is delivered, there's a certain amount of time that organizations and people take to actually take action and execute a decision. That introduces **decision latency**.

These latencies mean that we are losing time, getting farther and farther away from the point at which the event occurred. The farther away an action is from the point of the triggered event, the more the value of that action diminishes. We must respond and respond fast to the events happening around us.

In today's high speed, highly connected world, the window of opportunity to respond to events is shrinking, so we need to make sure that we are able to react quickly and reduce the amount of value erosion.

We can reduce action time by reducing each phase of latency, by making the data available and ready for analysis as close to the event as possible, by making sure that information is delivered fast, and therefore, enabling a faster decision through human collaborative systems.

## Meeting the Challenges of Big Data

"The enterprise data warehouse is dead!" That was the title of a 2011 article that ran in *Business Computing World* in the UK. But that vision is short-sighted. Successful companies will wed old and new together, creating a synergistic, greater whole.

The enterprise data warehouse is not dead, but traditional approaches to managing data are dead. We live in a different world from the one that existed when the traditional relational database was first architected. Back then the database dealt with recording and storing transactional data only and reporting happened for decision support. It was high value, highly structured data—not the mounds of data that may or may not have value that we face today. It was an era when time was measured by the calendar, not the stopwatch.

Today we face a completely different world. We generate vast amounts of non-transactional data—whether documents, Facebook posts, tweets, or log information coming off of our phones, web servers, and other connected devices.

We no longer want to report just against operational activities, but we also want to analyze, explore, predict, visualize and inspect in ways never imagined by those early database engineers.

Back when databases were designed, memory was extremely expensive. Just one terabyte of RAM cost over $100 million dollars. Today, we can get

it for less than $5,000. Because memory was expensive, database engineers built a database architecture centered on disk.

The problem is, disk is just too slow. Reading 1 petabyte of data off a disk sequentially would take 58 days using the fastest hard disk available today (according to Tom's Hardware website http://www.tomshardware.com/). SSD speeds that up to 2 days using the fastest SSD RAID, but the price is hefty: 1 petabyte of SSD RAID disk costs $12.5 million dollars.

That's why innovators have been finding new ways to store and process data, all in an effort to get around the disk bottleneck and improve response time.

### Distributed Databases

Distributed computing moved around the disk bottleneck by spreading the data across many disks that can be read simultaneously. In a perfectly balanced environment, a distributed database would have an equal amount of data across each machine. As a result, the maximum time to read the data would be a fraction of the time of a database stored on a single disk. For instance, if we split 1 Petabyte of data evenly across 10 disks that are read simultaneously, then the response time would be theoretically one tenth of the time of a single disk. Of course, in practice there is a cost to moving the data between the machines and coordinating a single result back to the user, but overall a database distributed across multiple disks can reduce response time.

### Enter Hadoop

Hadoop builds on the concept of distributed computing but opens up the platform to handle arbitrary data sets that do not necessarily follow a predefined schema and to analyze that data with any arbitrarily designed algorithm. This flexibility comes at a cost of course, such as the need for specialized programming skills. However, the Hadoop project has been evolving over the years to include subprojects that move beyond Hadoop Distributed File System (HDFS) and MapReduce.

Hadoop was originally developed at big Internet companies as a flexible tool to process Web logs. Based on its heritage, the original Hadoop HDFS and MapReduce projects made different assumptions than relational databases about how data is processed. In particular, the early Hadoop projects assume you want to read all (or at least most of) the data stored on your disks, which is why the MapReduce framework is designed to look for a predefined pattern within all of the data stored in HDFS. Furthermore, MapReduce algorithms are coded in Java or C/C++ in order to give the programmer the flexibility to define the search pattern as well as the schema of the result set. This combined capability ensured that the original Web companies could store any or all of the Web logs without having to do a lot of costly preprocessing of the data typically done with 'enterprise data.' Furthermore, as business analysts at the firms had a new idea for the fast evolving business, they could easily run a program to search for a new pattern. This flexibility meant that MapReduce queries usually took time to execute, forcing many companies to run them as a batch process.

### Columnar Databases

Moving database architectures from row-oriented storage models to columnar storage models helped to reduce the amount of data accessed on a single disk. This is fundamentally different than the original Hadoop project, which assumed the user wanted to read all of the data on a disk. The columnar database architecture assumes that any given query will need to read only a subset of the data on a disk.

The columnar database architecture assumes that the user typically will only want to access a small number of the attributes or columns within a database table. Imagine you have a table storing historical sales transactions with 8 columns: Year, Quarter, Country, State, Sales Representative, Customer, Product, Revenue. At the end of the year each department may ask different questions. For example:

- Finance: What was total revenue by year and quarter for last 3 years?

- Marketing: What was total revenue by product and by country?
- Sales: What was total revenue by sales representative?

In each case, the user is only accessing a subset of the columns. While this is a simplistic example, in practice many of the questions that people ask use only a small subset of the sometimes hundreds of columns in a table.

A columnar database stores all of the data associated with a particular attribute or column in the same physical space on the disk. In this way, when only 3 of the 8 columns of data are needed to answer a question, the database only needs to read 3 segments of the database from the disk instead of the entire thing.

Furthermore, by storing all of the data for a given column together, columnar databases can exploit the repeating patterns within a column's data in order to highly compress it, further reducing the number of bits read off disk. Consider the Country column from our example above. Storing the name "United States" as text would take at least 13–26 bytes of data depending on the encoding used. There are less than 256 countries in the world, which means that each country can be uniquely identified by using only 1 byte (8 bits) of information. So 'United States' could be replaced with, say, the number '1' compressing the column entry from 13–26 Bytes into 1 Byte. This form of compression is called tokenization.

It is very common for the rows to have a lot of repeated information. Building on our example, imagine that the 'country' column contains 'United States' for first 15 rows of the table, which has been replaced with the number '1' stored in a single byte in each row. This essentially means we have 15 entries in a row, each containing the number '1'. On disk then it looks like this '111111111111111'. This duplication can be replaced with the value, '1', and a count of the number of duplicate entries — something conceptually like this: '1D15', which says number '1' duplicated 15 times. This form of compression is called run-length encoding.

In summary, then, our first 15 rows of the 'country' column gets compressed from 195–390 bytes down to potentially 3 or 4 Bytes. Compression is important because it reduces the amount of data that gets read from disk. In our example above, reading 4 bytes from disk represents 200 bytes stored in other databases, which dramatically accelerates response time.

In summary, the data storage architecture organized around columns, which reduces the amount of data that needs to be scanned and also makes it easy to compress the data, makes columnar databases ideally suited for BI and analytic workloads.

### In-Memory Databases

In-memory databases take response times to a whole new level. They remove the disk from the equation for data access altogether, and only use it for logging and backup. In-memory databases leverage the power of today's processors to read and analyze data 1,000 times faster than reading data off disk.

Combine columnar data stores with in memory to highly compress the data, and soon you can see performance gains of 1,000, 10,000 … and in some cases customers have experienced results 100,000 times faster.

In-memory is the future of data management, and so the real-time SAP HANA platform for big data platform has SAP HANA at its core. Nonetheless, technology like Hadoop has a critical, complementary role to play. A complete big data solution is end-to-end in nature. It handles everything from low-level data ingestion, storage, processing, visualization, and engagement to analytic solutions and applications.

A complete big data solution has another characteristic as well: it handles all kinds of data. Location aware applications and applications that support mapping have made spatial data more important than ever, both operationally and in targeting customers. SAP HANA Spatial Processing helps process this important type of big data.

So much of big data is text, and text analytics, whether sentiment analysis of social media data or analysis of doctor's notes to help drive better healthcare, is

another key to big data. Text analysis is another key capability of SAP HANA (see the text analytics webinar to learn more).

Predictive analytics, largely the province of the data scientist, is another feature supported in SAP HANA through the Predictive Analytic Libraries or PAL.

Streaming analytics is another key area supported by SAP HANA and Sybase Event Stream Processor (ESP). Analytics where the velocity of the data is especially critical, such as in financial services, as well as in contexts like manufacturing where machine or sensor data is used and analyzed, an area referred to as the Internet of Things.

## Making Big Data Real

Since talking about big data in the abstract can't provide a clear vision of its benefits, this section offers concrete examples related to BI, data science, and real-time insights.

### Business Intelligence

Big data technology is needed not only for the many new types of data, but for large scale data warehouses. Case in point: the largest data warehouse in the world, as attested by the http://www.guinnessworldrecords.com/world-records/5000/largest-data-warehouse, holding 12.1 petabytes of data.

To explore another example, consider ARI, the largest fleet management services company in the world. In conjunction with its partners, ARI accounts for more than 2 million vehicles worldwide.

Maintenance management for the entire lifecycle of a single vehicle can involve more than 14,000 data points, including everything from information on minor repairs to regular preventive maintenance information and manufacturer updates and recalls.

ARI's data warehouse was straining under the load of the data. Its in-house ETL solution could not keep up with the growth in data, and analysis was

taking far too long. After a proof-of-concept, ARI migrated its data warehouse to SAP HANA.

ARI is able to perform deeper data analysis in less than four seconds (previously a manual process that took over 24 hours). The company also increased efficiency in call centers and improved first-time call resolution, resulting in higher customer satisfaction. "We have a 360-degree view of the data with our SAP solution," says Steve Haindl, EVP Technology and Innovation at ARI. "We can see what's working, where the opportunities are, and what customers no longer need. We can also tailor conversations about requirements to the interested party: CEO, fleet manager, or mechanic. All of this helps us to drive revenue, but most important, it helps us to keep our customers happy."

As ARI's Director of Information Management, Bill Powell explained, "There was a sea of information coming in and it could take up to two days to pull together, which affected our service levels. In-memory HANA means we can answer questions in seconds." ARI's Keith Allen added, "Our goal has been to drive greater efficiencies. The business can ask questions [of the data] and get responses directly. Customers can also build their own dashboards. It is self-service BI."

Many companies are considering moving their data warehouses and BI initiatives to SAP HANA and SAP IQ to speed up their analytic capabilities and drive faster value from their data.

### Data Science and Advanced Analytics

Healthcare is one of the most exciting big data stories there is, as seemingly intractable problems are beginning to find solutions through genomic analysis.

Invasive cancer is the second leading cause of death in the US. In 2008, approximately 7.6 million people died of cancer worldwide. The path to more successful cancer treatment lies in human DNA. More and more physicians are not only searching for changes in human tissue that signal cancer but are increasingly interested in the alterations of the human genome itself.

Mitsui Knowledge Industries (MKI) is working on http://www.saphana.com/docs/DOC-1799. They begin by pre-processing DNA sequences from normal cells and comparing them with cancer cells. Processing is done against large volumes of data. This pre-processing is run against data in Hadoop clusters and can take anywhere from several days to a week.

Next, they move relevant data into SAP HANA, where they perform complex analytical processes to identify variants from the pre-processed sequences. They also analyze what medicines might work against the mutated genes.

With SAP HANA, they take advantage of built-in predictive algorithm libraries (PAL) and integration with the open source R statistical tool to create predictive models to assess best treatment options for the patient.

Initially, MKI was using only Hadoop and R for analysis, but decided to add SAP HANA to reduce processing time so that they could deliver personalized results more quickly. Imagine being a patient in a doctor's office — being told that you have cancer — and that you have to wait for days before a treatment plan can be set. Now, imagine how you would feel if a customized treatment plan were provided to you the same day. This is MKI's goal — to provide personalized treatments to patients as quickly as possible.

MKI still uses Hadoop to pre-process large volumes of DNA (normal and cancerous) so that they have a strong foundation of existing sequences. But they now use SAP HANA to analyze a particular patient's DNA against related sequences from Hadoop to better predict the best medicines and treatment for the patient.

Hadoop is used to align the patient's DNA sequence with the normal sequence, because the data is in a semi-structured format, can be parallelized across multiple machines. Also, the MKI team is able to use an open source package for aligning genomes.

Identifying the mutations and predicting the best treatment requires a lot of highly iterative analysis. This is ideally done in SAP HANA. As a result MKI has been able to accelerate the overall time from 2 to 3 days to 20 minutes.

Furthermore, MKI believes they can get it under 10 minutes when they deploy a 64 node Hadoop cluster and a 40-core HANA machine.

### Real-time Insights

Arguably, big data can be most important where you need to analyze real-time data as it streams. Let's look at one such use of big data. McLaren Formula One cars run up to 350 km per hour on powerful V8 engines. These cars are loaded with sensors. About 120 sensors transmit data every second and some transmit even more frequently. Organizations like McLaren collect all this data to analyze it, and the pit crew uses it to make real-time decisions. Sensors provide information about wheel alignment, tire pressure, suspension, and so on and all of these parameters play a critical role in winning or losing a race. McLaren Applied Technologies leverages what it learns from this technology to drive other innovations, such as improvements in air traffic control and monitoring the vital signs of professional athletes.

These are just a few examples. Businesses can use big data to gain a 360-degree view of the customer by combining enterprise data with customer sentiment gleaned from social networks, customer service interactions, and web click-stream data. Service providers can proactively reach out to customers and keep them satisfied, loyal, and coming back for more.

## Who Uses Big Data?

A big data platform should meet the needs of all your stakeholders, from BI and analytic professionals to data scientists, to IT staff who help bring actionable insights to executive leadership, middle managers, and frontline workers, sometimes by even embedding those insights directly into business processes.

It is helpful to classify the different users into essentially three categories shown in Figure 2.

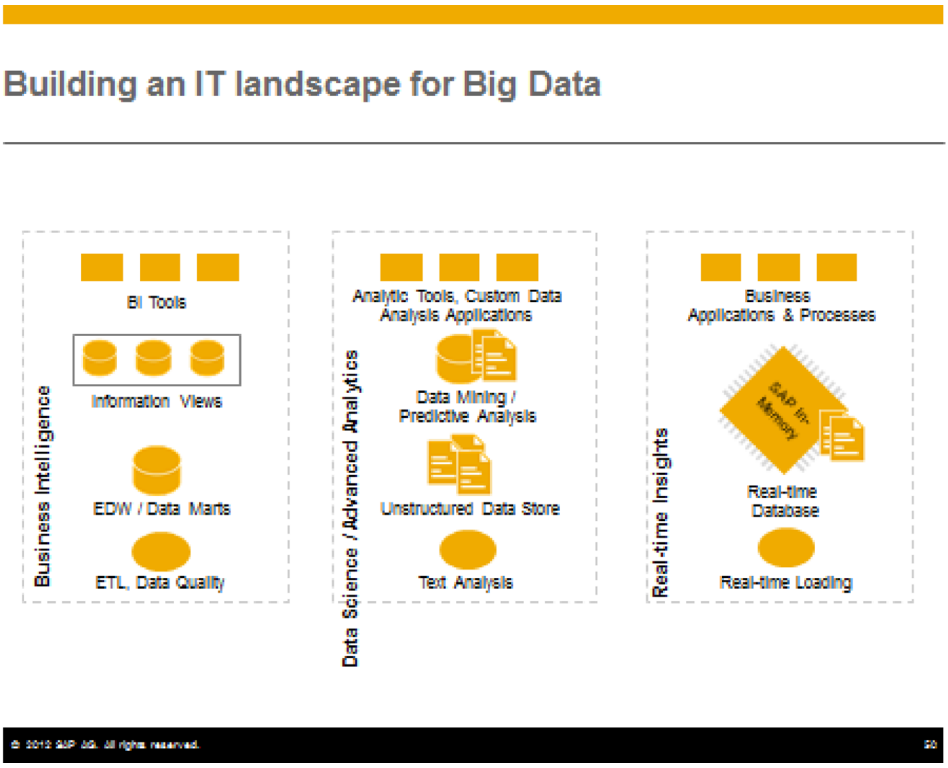



Figure 2. An IT landscape for big data, broken down by role

## Business Analyst

The business analyst provides the organization with precise, repeatable, accurate reporting on the data stored within the organization. They are supporting business operational decisions; it's all about the statements of fact, answered instantly. The business analyst is focused on reporting on the one and only truth. Normally, the information analyzed comes from data generated in transactional systems, data that is highly structured. There is an emphasis on data quality so that standardized reports can be executed with confidence that all of the numbers will line up. The information views of the

data have well understood meanings and there is a focus on unambiguous determinations.

The Business Analyst is in essence looking for an enterprise data warehouse and the rigor that entails. However, current data warehouses don't handle large datasets extremely well. As a result, many data warehouses contain summary data with much of the detailed information thrown away or stored in a highly complex BI landscape with many data marts, data caches, and generally many layers of technology. With SAP HANA, users can store all of the data without causing query response times to grind to a halt. This also makes it possible to remove the many data marts and data caches originally put in place to compensate for poor performance, greatly simplifying the data warehouse.

In some cases Business Analysts may benefit from access to Hadoop environments. If the data in Hadoop needs to be reported on, you may want to bring it into your enterprise data warehouse. Otherwise, it should be carefully structured and stored in Hadoop. Here's the key. The BI analyst uses GUI-based tools to access information and to generate reports. This requires data to be organized and structured in order to make it easily accessible and generated using forms. Projects like Hive and Pig help to do that for your Hadoop environment.

SAP BusinessObjects BI can now access Hadoop environments through Hive. With Hive, you define table structures to data stored in Hadoop. BI analysts can use the SAP BusinessObjects BI tools to create reports, dashboards, and explore data all inside Hadoop. BusinessObjects translates the users' actions into HiveQL commands, a language modeled after SQL. What is particularly powerful with BusinessObjects is that if the BI administrator has created the right universes, or access layers, the BI analyst can query data across various systems. In other words, you can take data from Hadoop and combine it with data from other data sources.

SAP IQ is an important platform to support the BI analyst. IQ is a disk-based bulk data store optimized for analytics. It can be used along with or even in place of Hadoop.

Of course, a critical step to providing the BI analyst access to Hadoop is to define what tables, columns, and so on are accessible and the relationships between them. SAP BusinessObjects BI gives the administrator the tools needed to do just that, including for a Hive implementation.

Here's the key: BI analysts need carefully controlled, structured access to Hadoop environments from their GUI tools.

### Data Scientists

In contrast, data scientists work at the other end of the information certainty spectrum. They deal with the uncertainty inherent in any large, complex organization and seek to draw conclusions that are statistically relevant but not completely certain. One example is predictive analytics, where large amounts of data are fed into models in order to predict what the future may hold. The data scientist may create custom systems to explore and probe the corporate data store and must be equipped with tools that interpret unstructured data and make sense of it for the organization and the problem domain.

The data scientist therefore requires as much flexibility as possible. The business analyst is skilled at using BI tools and understanding how the data applies to the business while the data scientist usually has very technical skills. Data scientists typically decide which tool to use based on the data that offers the most promise. They may choose a data mining technique, or techniques, and then select the tools that support the technique, such as the R statistical language, which is supported in SAP HANA.

While the BI analyst needs structure in a controlled environment, the data scientist wants a lot of freedom and flexibility. Depending on the analysis performed, they want to be able to run their algorithms in Hadoop using MapReduce algorithms, in-memory, or in the database using in-database analytic algorithms.

### Operational Users

Operational users are involved in the day-to-day operation of core business processes. They are the frontline workers such as call center operators, marketing campaign managers, warehouse personnel, and sales representatives. Operational users can benefit from information that helps them make decisions in the moment, often based on insights uncovered by business analysts and data scientists. This information is often delivered in the form of dashboards, daily reports, or even predictive models embedded in enterprise applications. The challenge of real-time analysis is to feed automated insights back into the decision loop fast enough to guide the action of the human or the machine making crucial decisions.

Operational users typically are not technical nor do they have experience in using analytic and reporting tools. In essence, the solution requires development of user interfaces suited to how operational users need to consume the information.

## SAP HANA: The Heart of SAP's Big Data Ecosystem

Like most things in the real world, the big data landscape is complex. Competitive advantage doesn't come from having one tool, but from having the right toolset to support business needs.

SAP provides an integrated set of data management solutions for big data: HANA for real-time analytics on operational and transactional data, SAP IQ for petabyte scale storage and analytics of less time critical data, and Hadoop as a massive data lake where data can be stored and explored. SAP does not market its own Hadoop distribution, but provides an open platform to work with a variety of Hadoop distributions. Lastly, SAP provides a suite of Information Management solutions to integrate systems, ensure data quality, and manage the overall data landscape.

The heart of SAP's big data platform is SAP HANA. At its core SAP HANA provides an in-memory, columnar, distributed database architecture designed to handle massive datasets. Since the SAP HANA database resides entirely

in-memory all the time, additional complex calculations, functions, and data-intensive operations can happen on the data directly in the database, without requiring time-consuming and costly aggregations.

SAP IQ was the first commercial column store, which is designed to scale to petabyte scale database size. While it is not in-memory like SAP HANA, it has excellent performance characteristics with a rich SQL layer, patented indexing, and a disk-backed store. SAP IQ and SAP HANA are integrated to work well together through smart data access, which allows remote tables to be queried as though they were local tables. This provides real-time analytics along with data scalability. Smart data access in effect creates a logical data warehouse.

In essence, SAP HANA smart data access enables the creation of a logical data warehouse, where data in HANA, IQ, and Hadoop can be mapped at a higher level, freeing the analyst from understanding exactly where in the landscape data resides. This solution amplifies the value of big data across your data fabric by enabling working with data sets stored in a variety of places including Hadoop. For more information, see the Data Virtualization webinar.

SAP HANA can access data in other data sources such as Hadoop to extend the reach of its processing power. Hadoop provides vast and flexible storage for data objects, independent of their structure and size. Hadoop is perfectly positioned to store the very large data sets that are too big to fit into memory and that require a preprocessing step before they can be easily analyzed. By connecting SAP HANA to Hadoop, you can run jobs in Hadoop that load information into HANA and then provide super-fast final analysis, as described in the section on personalized cancer treatment earlier in this chapter.

When you put SAP HANA, SAP IQ, and Hadoop together, you have three data processing domains with different strengths that combine to form a big data processing backbone. Together, these three components provide real-time capabilities along with extreme scale. Data can be processed with the appropriate technology depending on its characteristics — hot data in HANA,

warm data in IQ, and a vast data lake in Hadoop — where data can be stored, processed, and aggregated without constraints on size, format, or cleanliness.

SAP Data Services is a sophisticated ETL and text processing tool, and ESP can capture streaming sources of machine generated data. SAP BW is a rich data warehouse layer on top of SAP HANA. SAP BusinessObjects BI universes can pull Hadoop and database sources together to serve up information to business applications. All of this technology works together to bring big data into the enterprise (see Figure 3).

SAP® Event Stream Processor (SAP ESP) is a mature and high throughput complex event processing engine that allows for integration of real-time data streams into the big data environment. It is a key tool for building real-time applications that help to formulate a response to real-time data.

Figure 3 shows the SAP HANA architecture in the context of big data.

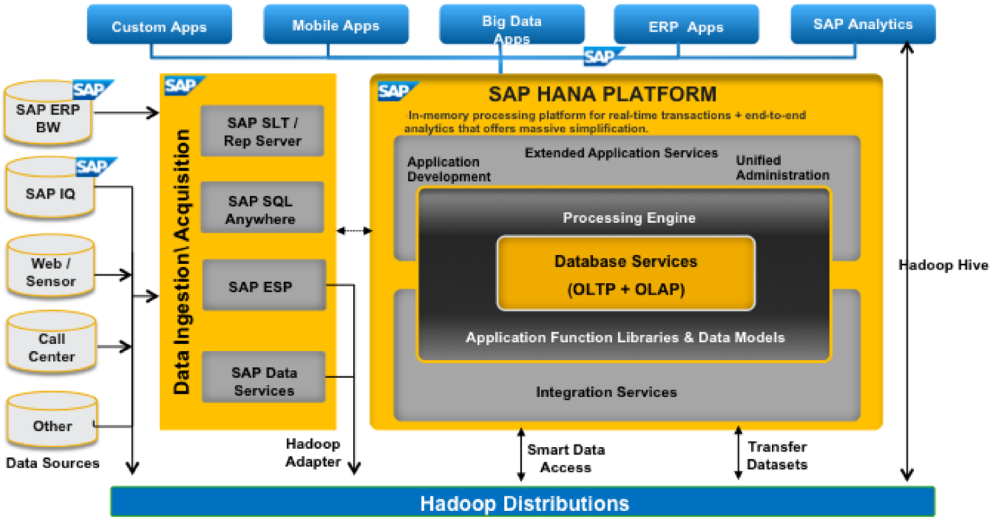
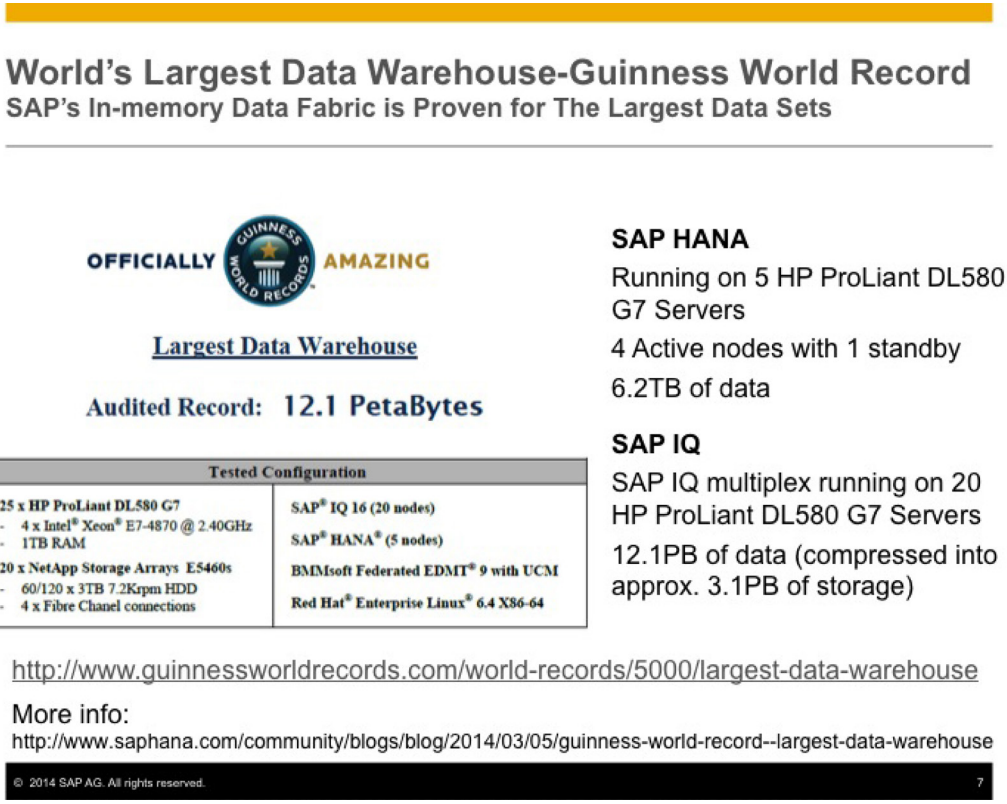



Figure 3. The SAP HANA Platform

## The World's Biggest, Big Data Database

In early 2014, SAP generated a new world record for the world's largest data warehouse using the SAP HANA® platform and SAP® IQ software. This independently audited 12.1PB data warehouse has been recognized by Guinness World Records, and is four times larger than the prior record.




This new world record demonstrates the ability of SAP HANA and SAP IQ to efficiently handle extreme-scale enterprise data warehouse and Big Data analytics. SAP and its partners had previously set a world record for loading and indexing Big Data at 34.3 Terabytes per hour.

## Conclusion

SAP's big data technology simplifies the IT landscape. SAP HANA provides speed for dealing with big data in real-time. It can also speed up traditional enterprise data warehouse applications, putting them on steroids. This chapter has touched on many points that are deserving of their own chapters, and we hope you will explore the links to learn more as well as checking out SAPHANA.com. The complementary nature of SAP HANA, SAP IQ, and Hadoop supports every big data use case, whether it's driven by BI analysts, data scientists, or IT seeking to help big data inform the real-time enterprise.