NS2™

SAP NATIONAL SECURITY SERVICES

CIO Guide to Using the SAP HANA® Platform for Big Data

At SAP NS2, we do one thing: support the mission of national security. We leverage the most advanced commercial technologies to protect the lives, assets and information of American citizens. Our expertise includes innovative computing and analytics solutions, cloud technologies and secure support, all delivered by U.S. citizens on U.S. soil.

DRIVEN TO DEFEND

CIO Guide to Using the SAP HANA® Platform for Big Data

February 2016

Table of Contents

The Best-Run Businesses Run SAP®

Executive Summary

This guide supports CIOs in setting up a system infrastructure for their business that can get the best out of Big Data. We describe what the SAP HANA® platform can do and how it integrates with Hadoop and related technologies. We examine typical data flows where different technologies interplay, looking at data lifecycle management and data streaming. Concrete use cases point out the requirements associated with Big Data as well as the opportunities it offers, and how companies are already taking advantage of them.

Big Data is often characterized by the three Vs: volume, velocity, and variety. These characteristics pose great challenges for conventional disk-based relational databases. The different categories of data require different storage capabilities and performance, which involve different costs.

The SAP HANA platform offers several types of data storage and processing engines. Online transaction processing (OLTP) and online analytical processing (OLAP) applications can now easily run in one system, on one database. In-memory stores in SAP HANA are recommended for high-value data (hot data) that must be accessed and processed with extreme speed, for data that is frequently changed, and when you need the platform's native features. Customers typically use in-memory stores for (compressed) data volumes up to several terabytes.[1]

The dynamic tiering option extends the SAP HANA database with disk-based columnar tables, based on SAP® IQ software technology. This option is recommended for storing big sets of structured data, when high OLAP performance and deep integration into SAP HANA is important, and when the processing features of Structured Query Language (SQL) are sufficient. Dynamic tiering can be used, for example, for lower-value data for which in-memory performance is not required (warm data). It can manage data volumes from several hundred terabytes to petabytes.

Hadoop is suited for raw data that can grow infinitely, for unstructured and semistructured data, and when massive scale-out is required for processing. With Hadoop you can flexibly scale out with low initial cost. Hadoop is also suited for data from business applications that is no longer required (cold data).

The SAP HANA Vora™ engine is the recommended SQL engine for high-performance analytics on structured data in Hadoop. It enables optimized access to data in SAP HANA from Hadoop or Spark.

For all data, SAP HANA serves as the central point for data access, data modeling, and system administration. Due to its openness, the SAP HANA platform can be extended with non-SAP technology depending on the requirements. This flexibility makes the platform a sustainable investment. By streamlining system administration and software lifecycle management, SAP HANA enables CIOs to simplify their system landscape and significantly reduce cost of ownership.

1. In SAP HANA, a volume on the order of terabytes results after compression. Therefore, this actually equates to much larger data volumes in conventional systems.

Introduction

Technology has the ability to shape our world. Big Data is one of the most important technology trends that will impact our world between now and 2020.

Arguably, Big Data is an artificial category created by technology vendors as a convenient way to reference certain new tools (the value of which is undeniable). The term is broadly used to refer to large or complex data sets that traditional data processing applications are not able to manage. In the past five years we have created more data than all prior years combined, and having all this new data is making business operations much more complex.

In addition, a lot of data does not automatically equal a lot of useful information. An effective Big Data infrastructure should be able to separate the background noise from the valuable signals that can be translated to actionable insights.

There are many choices when it comes to designing and setting up a suitable Big Data system infrastructure, but there has been little guidance regarding the best approach for both exploiting the potential of Big Data and meeting enterprise-specific requirements. This document provides orientation for CIOs so they can choose the right storage and processing options for a given Big Data application and understand the impact and consequences of their decision. It helps to answer questions about what type of data should be stored in in-memory tables, in dynamic tiering, and in Hadoop, as well as which processing engine should be used for a given task.

Making the wrong decision can lead to unnecessarily complex and expensive solutions that do not meet the business requirements of Big Data. Consider, for example, an application that uses SAP HANA to store its business data and has to manage additional data coming from devices. Assume that the device data is structured, the total expected volume is in the gigabytes range, and the data is processed with SQL. Since such data can be efficiently managed in the SAP HANA database, a solution architecture that requires an additional Hadoop cluster would add unnecessary complexity and cost of operations.

The decision for a particular storage option or processing engine depends on several dimensions, including performance, data volume, cost, features, and operational readiness. This guide helps CIOs understand these dimensions and make the right decision for their business.

The guide is structured in the following sections:

- The "Challenges and Opportunities of Big Data" section describes briefly what makes Big Data challenging yet, at the same time, a source of benefit for today's businesses.
- "Big Data Reference Architecture" explains the lambda architecture, which is one of the current reference architectures for Big Data.
- "SAP HANA Platform for Handling Big Data" gives an overview of the SAP HANA platform and its options for managing Big Data.
- "Big Data Scenarios and Data Flows" explains typical Big Data scenarios and data flows and how they influence the Big Data infrastructure setup.
- "Big Data Use Cases" describes two real-world use cases.
- "SAP HANA Platform: Open, Flexible, Integrated, Scalable, and More" summarizes the key characteristics of the SAP HANA platform that make it a sustainable investment.
- "Find Out More" provides links to further information.

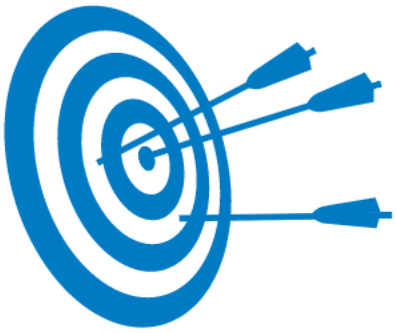
Challenges and Opportunities of Big Data

In today's networked economy, characterized by hyperconnectivity or instant connection to the business network, we can confidently anticipate that data quantities will continue to grow at high rates. At the same time, this data will be of different types – structured and unstructured, and from low value to high value.

In the last decade, companies like Google, Facebook, and Netflix have led the way in collecting and monetizing huge amounts of data generated by consumers' everyday activity. They look on this data as a strategic asset – every decision in their organizations is data driven, as is every product they sell. This has created enormous interest among traditional enterprises, which can easily see the benefit of putting their data to work in the same way.

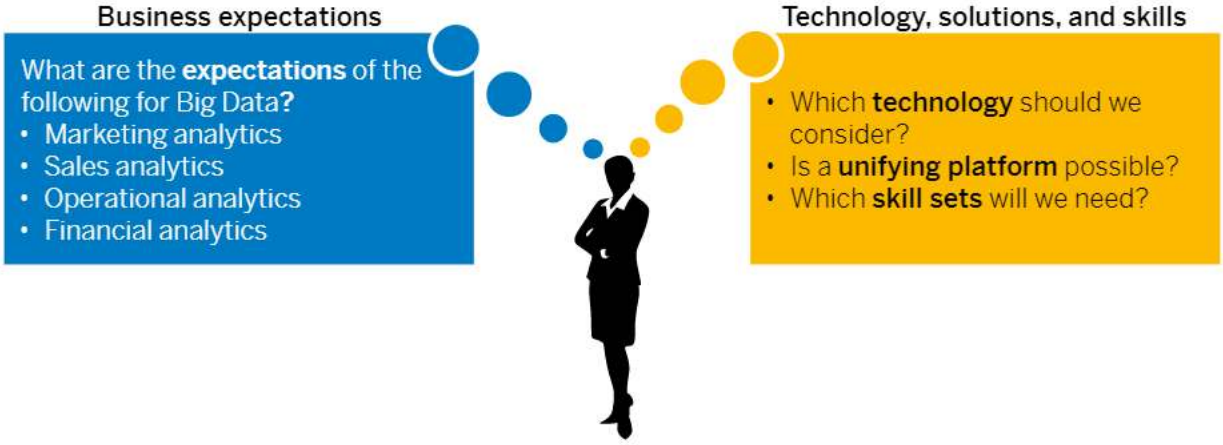
There has been an explosion of new technologies, new data types and sources, and new ways of using existing technology. From mobile and cloud computing to social media and sentiment to machine, log, and sensor data, data science is wringing value from massive amounts of structured and unstructured data, and high-velocity streaming data is providing insight and driving decisions like never before.

Possessing data does not add value in and of itself, but being able to use it to make timely, meaningful decisions that impact business is enormously valuable. Before Big Data can be monetized and turned into a strategic asset, however, today's CIOs need to consider several things. To start, the expectations of the business should be matched by the most suitable technology (see Figure 1).



Big Data is often characterized by the three Vs: volume, velocity, and variety. These characteristics pose great challenges for conventional disk-based relational databases.

Figure 1: Matching Business Expectations with Technology, Solutions, and Skills



The left side of Figure 1 shows common expectations regarding Big Data that require consideration of:

- New digital channels
- Possible correlations between transactional and analytics data from the enterprise with data from other sources (for example, weather data or social media) that might be meaningful
- The ability to create simulations and sophisticated data visualizations that depict insights from data in new, more compelling ways

On the technology side, CIOs should get familiar with the available technologies and their capabilities in order to make the right decisions to build a system infrastructure that satisfies business expectations without disregarding costs. For example, a unifying platform can reduce total cost of operations and at the same time preserve existing investments in technology, staff, and training.

To manage Big Data well, enterprises must have people with dedicated skills to exploit the opportunities that it offers. The main task of data analysts or data scientists is to separate the background noise from the meaningful signals to enable solid decision making and take appropriate actions. SAP HANA is already helping businesses to unlock this valuable information by addressing at least one very important aspect: the ability to perform real-time analytics on very large data sets. Not only data scientists but also managers and executives can now get insight into their current state of affairs at any time and at "the speed of thought."

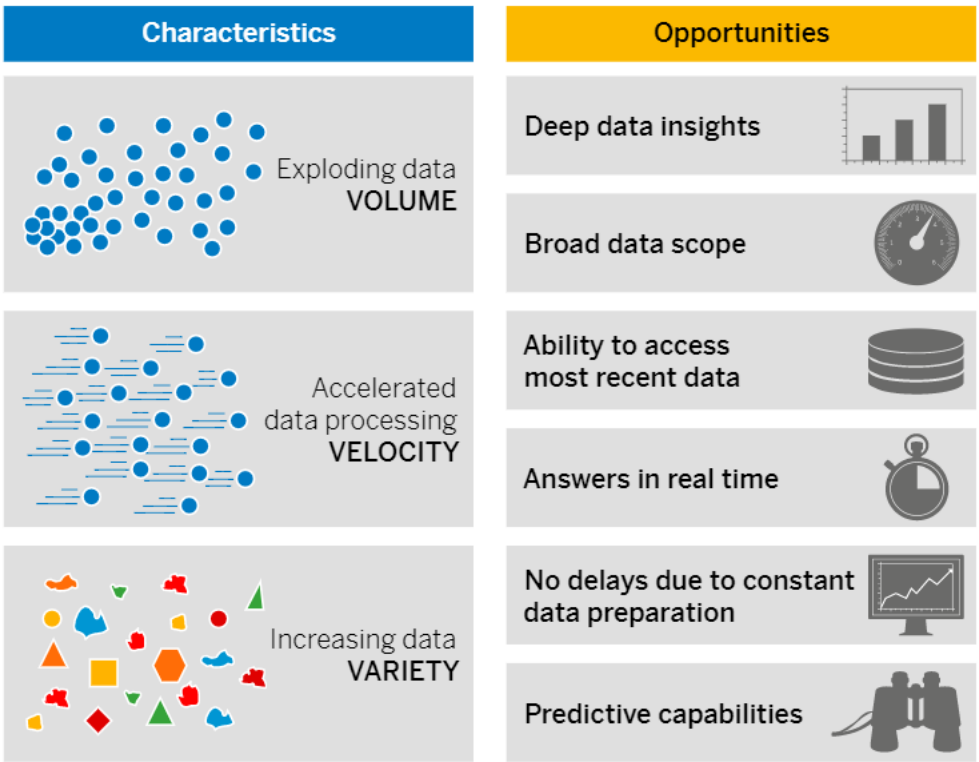
CHARACTERISTICS OF BIG DATA

The most conspicuous and technically measurable characteristics of Big Data are often referred to as the three Vs: volume, velocity, and variety (see Figure 2).[2]

Not only is the volume of data large, but it is arriving ever more rapidly (velocity) – consider machine data generated on the factory floor, or algorithmic trading data generated by financial markets. The data is also of many different types (variety) – from comments about products on Twitter or Facebook and logs of customer behavior on a company Web site to sensor and weather data.

A general architecture pattern has been postulated to account for exploding data volumes, fast-growing data quantities that require processing and storage, and increasing data variety. This pattern is called the lambda architecture, which we discuss next.

Figure 2: Characteristics and Opportunities of Big Data



2. "Value" and "veracity" are further dimensions that are considered to be characteristic Vs of Big Data. Value refers to the ROI the data is able to provide to the business. ROI accounts for both the potential benefit that can be derived from the data and the cost of storage media. Veracity refers to the ability to trust the data used to make decisions (which, ultimately, is required of all data).

Big Data Reference Architecture

The techniques that have been used to deal with Big Data as a strategic asset are based on technologies that allow the collection, storage, and processing of very large amounts of data at a low economic cost. Initially, some of these technologies (for example, Hadoop) were primarily used for batch workloads. However, in the last several years other technologies (for example, Spark) have emerged that enable both batch and real-time processing to work in parallel in the same infrastructure. It is precisely this combination of capabilities that is the foremost requirement of Big Data architectures. The lambda architecture describes how this requirement can be achieved. The Big Data community regards it as an important reference architecture for Big Data (although it has pros and cons). It is particularly well suited for predictive analytics where patterns are identified within a historical data set on a regular basis and incoming records are checked in real time to see if they correspond to these patterns.

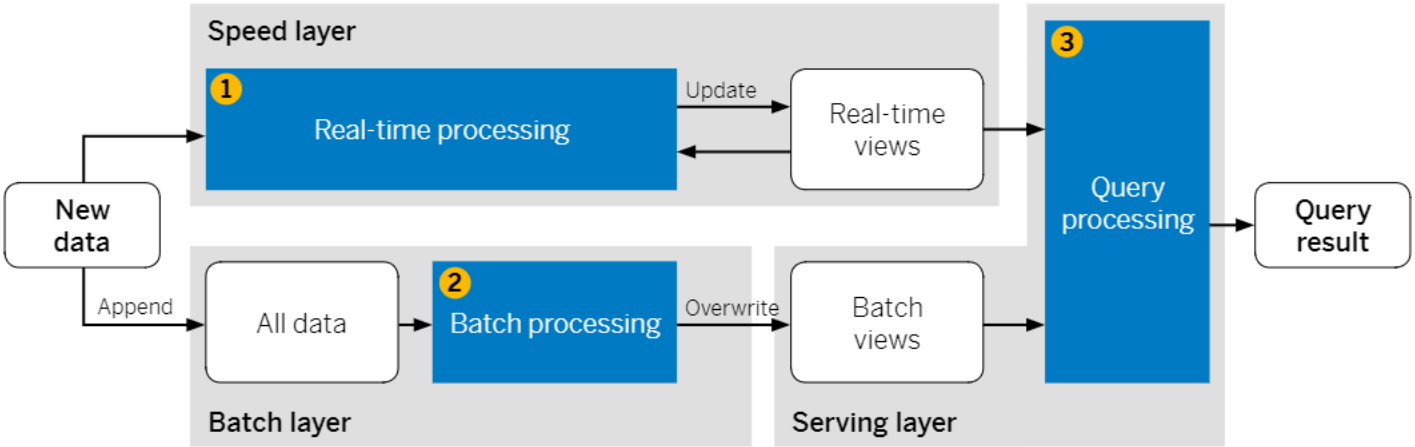
The most important characteristics of a lambda architecture are:

- Fault tolerance: ability to satisfy requirements in spite of failures (if a fault occurs, no information gets lost, since the incoming data can be recomputed from the append-only master data set)
- Scalability: flexibility to accommodate data growth
- Low latency for reads and writes: minimum time delay of system responses
- Real-time results: quick return of results despite data load

Lambda architecture is defined by three functions: real-time processing, batch processing, and query processing.

Figure 3 shows how these functions interact. Additionally, the master data set ("all data" in the figure) is being continuously updated by appending new data. To compensate for the latency of batch processing, three architectural layers are defined: the batch layer, the serving layer, and the speed layer.

Figure 3: Lambda Architecture



Incoming new data is routed into both the batch layer and the speed layer. In the batch layer, the incoming data is appended to the master data set – for example, in a Hadoop Distributed File System (HDFS). Batch jobs are used to read the master data set and produce precomputed and preaggregated results called batch views. The potentially long-running batch jobs are continuously reexecuted to overwrite the batch views with newer versions as soon as they are available. The batch views are loaded into a read-only data store in the serving layer, to support fast and random reads. Real-time views are not copied because an ad hoc creation of the view is desired; every query should be answered with the newest possible data. Therefore, the real-time view is part of the real-time layer and not reproduced in the query layer.

In the lambda architecture pattern, incoming queries are able to act on a combination of results from the batch and real-time views. Consequently, databases typically index the batch view to allow for ad hoc queries and have one component to merge both real-time and batch views.

Fault tolerance is achieved by maintaining an immutable, append-only master data set in which all incoming data is stored. If a fault occurs in any processing step, everything can be recomputed from the master data set. The continuously growing data is stored in HDFS, for example, as scalable and reliable distributed storage. The data in an HDFS environment is traditionally processed with MapReduce batch jobs.

For all data, SAP HANA serves as the central point for data access, data modeling, and system administration.

The serving layer enables fast queries in the batch views, but these views only contain information from the time before the producing batch jobs were started. To compensate for this time gap, all new data is fed into the speed layer in parallel. The speed layer processes new data in real time with stream-processing technology (such as Apache Storm or Spark Streaming). The results are used to incrementally update real-time views in a database supporting fast random reads and writes (for example, by using Apache Cassandra[3]).

The speed layer needs to approximate the results because it has only current information and no historical data. As explained, real-time views and batch views are merged in the serving layer when a query is executed.

Recently, the lambda architecture has been criticized because it creates a complex landscape with many different components that are loosely integrated, and it requires implementing different code bases on different technology stacks in the batch and speed layer. At the same time, the implemented functions must be maintained and kept synchronized so that they produce the same results from both layers. The effort involved is huge because of the complexity of the distributed systems.

But in spite of these criticisms, lambda architecture continues to be a valid way to handle Big Data. It is ultimately a matter of making lambda requirements more manageable.

SAP HANA enables you to implement a complete lambda architecture from preintegrated components of the platform, including the processing engines and data stores. SAP HANA even goes beyond lambda since it comes with options for integration with remote data sources and supports data integration and replication, data transformations, data quality operations, event streaming processing, and federated query execution. Moreover, SAP HANA offers the advantages of an integrated data platform – simpler installation, administration, lifecycle management, and development – because all three layers of the lambda architecture can run within one system and all persisted data can be accessed within the same database.

3. Apache Cassandra is an open-source distributed database management system initially developed at Facebook to deal with its in-box search feature. For details, see the DATASTAX Web site, "About Apache Cassandra."

SAP HANA Platform for Handling Big Data

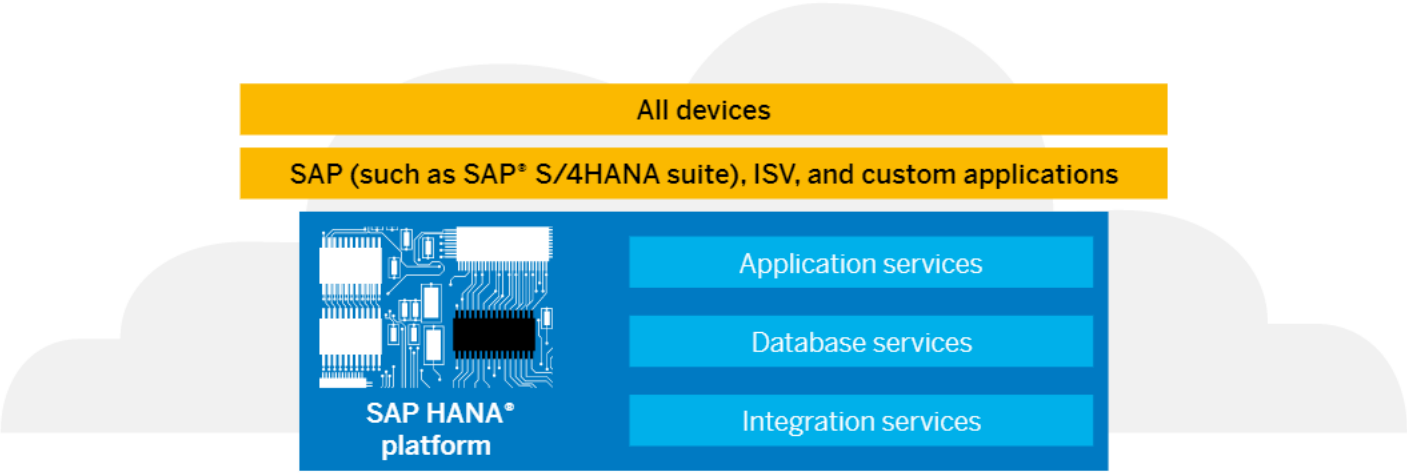
As mentioned earlier, more and more software solutions require capabilities to manage and process Big Data holistically, independent of whether the source is a machine sensor or social media. Big Data typically needs to be combined with traditional business data created by enterprise applications. SAP HANA is the SAP strategic platform for unifying and combining all this data. It is ideal for central data management for all applications because it is open and capable of handling not only transactional but also analytics workloads all on one platform. As described in this section, integration capabilities in SAP HANA make it possible to combine it with other technologies (such as Hadoop and members of its family) to obtain the most suitable and effective Big Data landscape.

The SAP HANA platform includes application services, database services, and integration services (see Figure 4). Since database and integration services are most relevant for Big Data infrastructures, we discuss them first.

DATABASE SERVICES

Database transactions are processed reliably because the system is compliant with ACID[4] and SQL standards. The database services are accessible through Java Database Connectivity (JDBC), Open Database Connectivity (ODBC), JavaScript Object Notation (JSON), and Open Data Protocol (OData). Ultimately, this means that SAP HANA is both standards based and open for connection through commonly used application programming interfaces (APIs) and protocols, which facilitates its adoption and adaptation to existing infrastructures.

Figure 4: Overview of the SAP HANA Platform



ISV = Independent software vendor

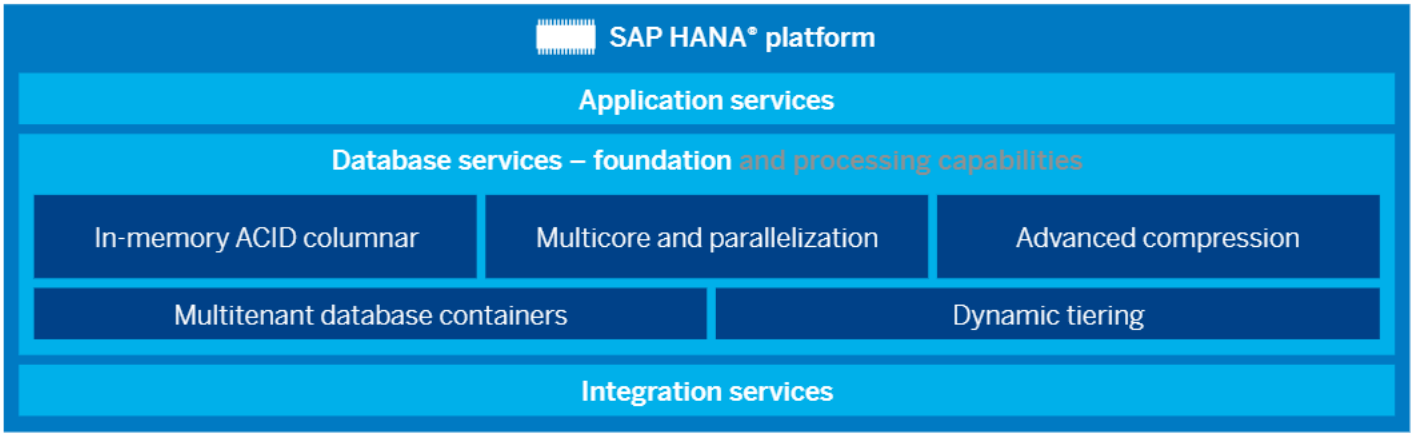
4. Atomicity, consistency, isolation, and durability.

The database services encompass foundation and processing capabilities. The foundation consists of functionality that turns data into real-time information, with no sophisticated tuning required for complex and ad hoc queries (see Figure 5). OLTP and OLAP can run on a single copy of data in the same system because data is in-memory, and the columnar store in SAP HANA can handle both types of workloads with high performance.

In mixed environments with both OLAP and OLTP operations, SAP HANA deals with a blend of hundreds to tens of thousands concurrent statements, from simple, short-lived, and high-priority transactions to deeply complex, long-running analytical queries that consume a lot of resources from the host. A dedicated workload manager helps ensure that this is done effectively by controlling parallel processing and prioritization of processing activities.

Because SAP HANA uses multiple tenant databases that can run in one instance of SAP HANA, it is cloud ready and allows secure, efficient management of a shared infrastructure. Moreover, SAP HANA maintains a strong separation of data, resources, and users among tenant databases. Multiple databases can be managed as a unit. Additionally, SAP HANA provides the flexibility to allocate memory or CPU to each tenant database.

Figure 5: Database Services – Foundation



ACID = Atomicity, consistency, isolation, and durability

Due to its advanced compression capabilities, SAP HANA can support scale-out deployment of several terabytes – it is not confined by the size of memory. Assuming a data compression factor of 7 and that approximately 50% of the memory is used for query processing, a 3 TB single node system, for example, could store a database with an uncompressed size of 10 TB. Scale-out systems with much more memory are also possible, and there exist productive scale-out systems for databases with an uncompressed size in the range of hundreds of terabytes. The largest certified scale-out hardware configuration for SAP HANA has 94 cluster nodes with 3 TB of memory each.[5] This extreme certified scale shows that SAP HANA is technically able to store Big Data in-memory, even though this is not often done in practice.

In addition, warm data can be stored on disk in a columnar format and accessed transparently. This is the dynamic tiering option, which extends SAP HANA with a disk-based columnar store, also called an extended store. It runs in the extended store server, which is an integration of SAP IQ into SAP HANA. The extended store server can manage up to petabytes of data and is optimized for fast execution of complex analytical queries on very big tables.

Possessing data does not add value in and of itself, but being able to use it to make timely, meaningful decisions that impact business is enormously valuable.

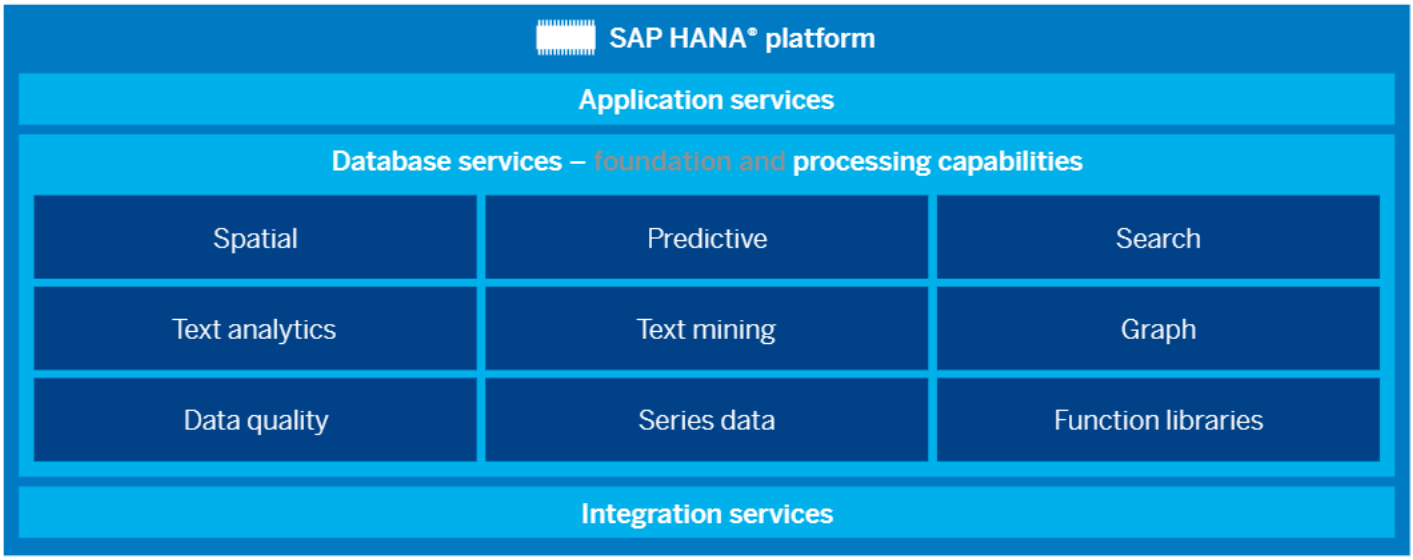
5. See the certified hardware directory for SAP HANA at http://global.sap.com/community/ebook/2014-09-02-hana-hardware/enEN/index.html.

The processing capabilities of the database services (see Figure 6) allow running applications with almost any data characteristics in the same system. These capabilities include:

- Analysis of sentiment by summarizing, classifying, and investigating text content
- The ability to search across structured and unstructured data
- Persistence, manipulation, and analysis of network relationships and graphs without data duplication
- Built-in business rules and functions that accelerate application development
- Prepackaged predictive algorithms that operate on current data, as well as open predictive and machine-learning abilities – for example, through the integration of an R server
- Interactive planning without the need to move data to the application server
- Cleansing of locally generated and imported data without the need for postprocessing.

Particularly in the context of the Internet of Things, the capabilities of SAP HANA facilitate handling of series data, mostly consisting of successive events collected over a predefined time interval.

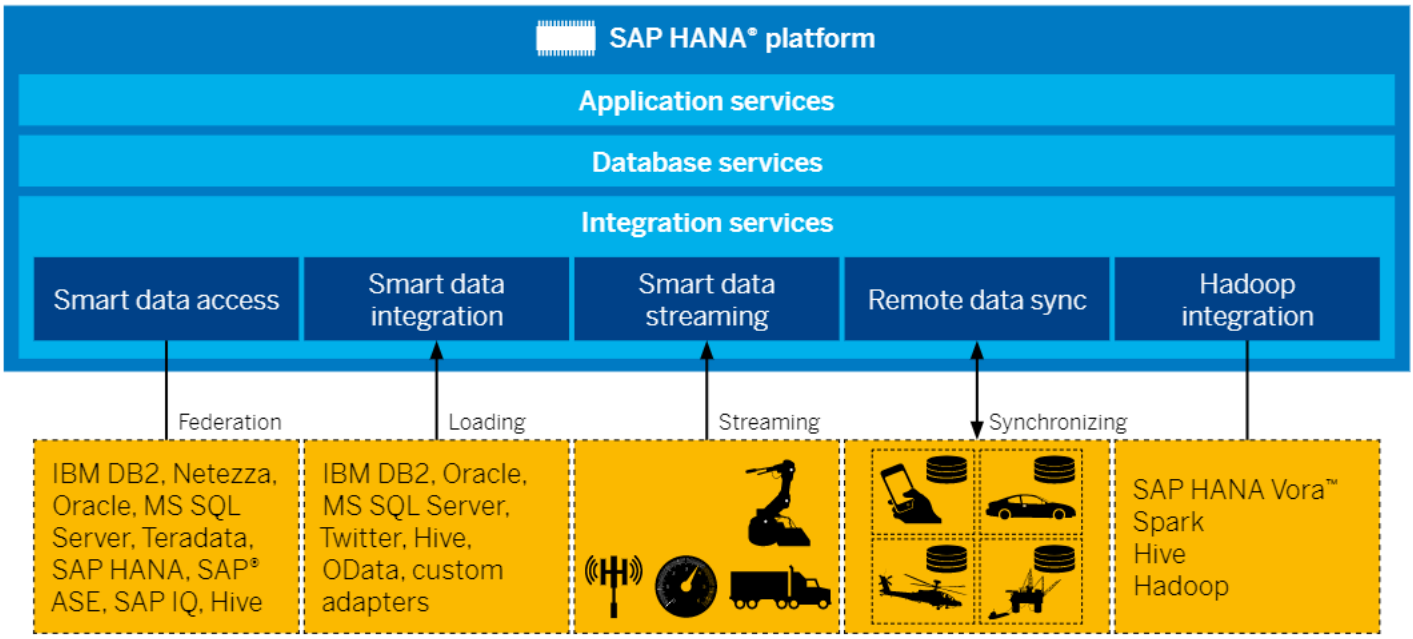
Figure 6: Database Services – Processing Capabilities



INTEGRATION SERVICES

Integration services play a strategic role in handling Big Data coming from any source and using it to provide a complete view of the business (see Figure 7).

Figure 7: Integration Services of the SAP HANA Platform



SAP ASE = SAP Adaptive Server Enterprise

The integration services of SAP HANA allow access to information in different data sources. These services enable replication and movement of almost any type of data in near-real time:

- Smart data access (SDA) enables remote query execution, also known as data virtualization. By means of virtual tables, data in remote systems is made available from queries executed in the SAP HANA platform.
- Smart data integration (SDI) can be used for both batch and real-time data provisioning from a variety of remote data sources. Aside from supporting additional kinds of data sources, SDI also includes an adapter software development kit (SDK) for customers and partners. SDI partially uses SDA – for example, it uses the concept of remote source systems and virtual tables. It also plugs into the SDA federation framework for remote query execution.
- Smart data streaming (SDS) enables capture and analysis of live data streams and routes them to the appropriate storage or dashboard. It includes a "streaming lite" deployment option that can run on small devices in the edge (for example, on Application Response Measurement [ARM]–based Linux). It can capture and preprocess data from sensors and machines and send results to a central core or cloud-based SDS.
- Remote data synchronization can be used for bidirectional data synchronization with embedded SAP SQL Anywhere® solutions and UltraLite databases running on devices. This is particularly indicated for data synchronization across high-latency or intermittent networks.
- Hadoop integration involves multiple access points from SAP HANA to Hadoop data (through Spark, Hive, HDFS, and MapReduce). Furthermore, SAP HANA Vora is an in-memory, column-based SQL engine that complements the SAP HANA platform with a native SAP query processing engine on Hadoop. It is designed for high-performance queries on Big Data volumes in large distributed clusters. SAP HANA Vora is integrated into the Spark computing framework.

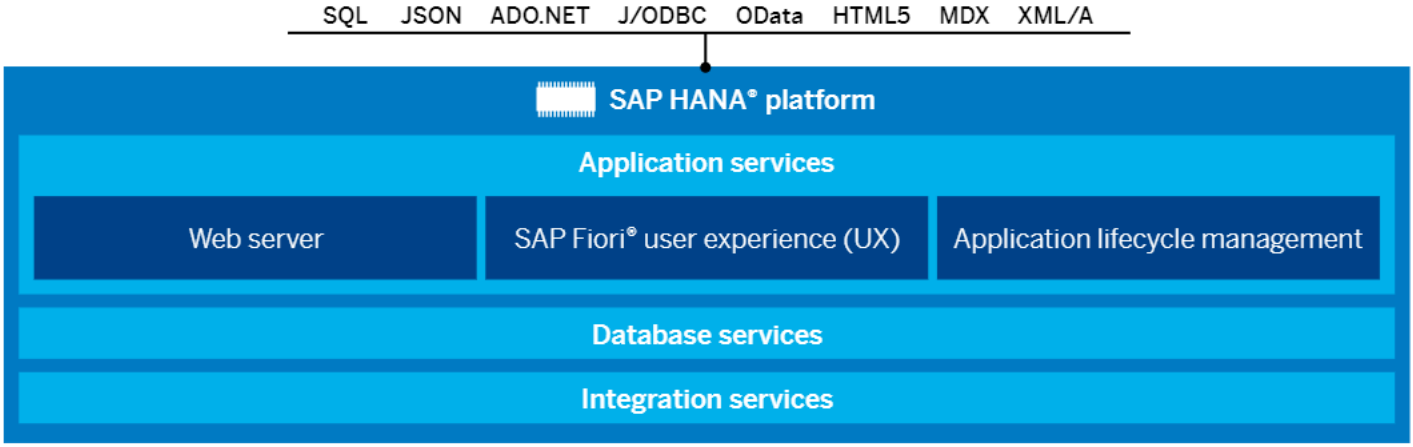
Along with the database services, the integration services enable the SAP HANA platform to handle Big Data regardless of its characteristics (volume, velocity, and variety). In addition, the services offer openness and connectivity with virtually any technology, and the most important tools and toolkits are already closely integrated into the SAP HANA platform environment.

APPLICATION SERVICES

Big Data is of limited value unless you can operationalize the insight through applications and business processes. New business processes may even be created by enhancing existing applications or developing new ones. The application services of SAP HANA support Big Data management by enabling all kinds of applications to create new or use existing data stored in the platform. Moreover, when building Big Data applications, these services help to simplify the landscape by running applications on the same platform. The application services deliver a first-class user experience on any device through the SAP Fiori® user experience (UX) technology. In addition, they support open development standards such as those for HTML5, JSON, and JavaScript. Built-in tools are included that support development, version control, bundling, transport, and installation of applications (see Figure 8).

The constituent services of the SAP HANA platform enable you to deal with the challenges and requirements associated with Big Data. The database foundation services support you when dealing with structured (OLTP and OLAP) and unstructured data from any source, and support high performance on one platform with different storage options. The processing capabilities facilitate handling data of almost any type, allowing sentiment and predictive analysis, among other types. The integration services facilitate connection to virtually any data source and cover the requirements of lambda and other architecture principles that may be related to Big Data handling. And finally, the application services provide a state-of-the-art look and feel for all applications running on SAP HANA and help simplify system landscapes by letting you run applications on the same platform.

Figure 8: Application Services of the SAP HANA Platform



SQL = Structured Query Language
JSON = JavaScript Object Notation
ODBC = Open Database Connectivity
OData = Open Data Protocol
MDX = Multidimensional Expressions

Big Data Scenarios and Data Flows

Having introduced the elements of the SAP HANA platform, it is now time to show how they interact and enable you to deal best with your Big Data requirements. In this section we examine typical Big Data scenarios and the associated data flows. This will help you better understand the characteristics of the different storage and processing options. We will discuss two main scenarios:

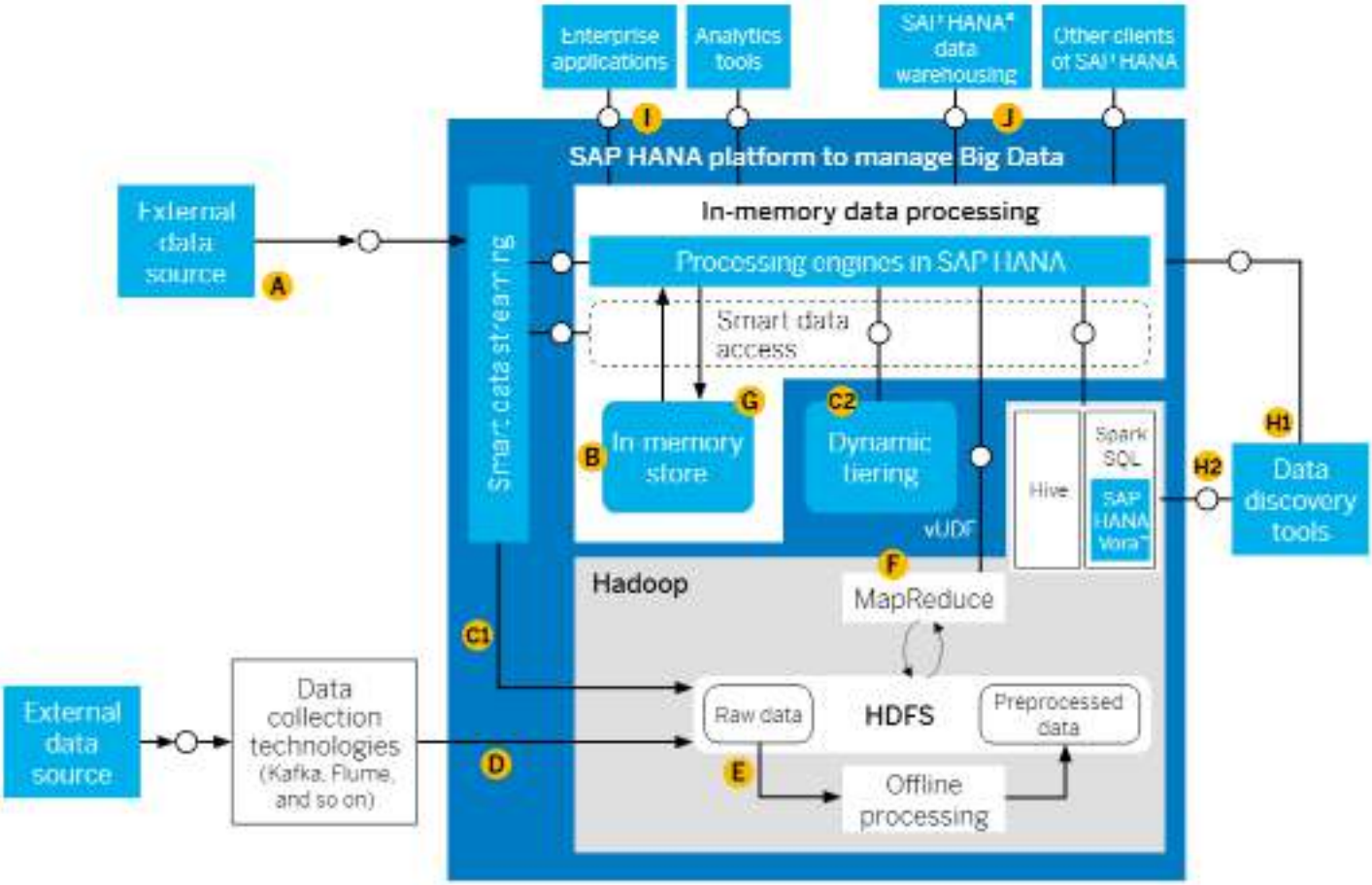
- How incoming external data is stored and processed
- How the lifecycle for data created by enterprise applications is handled (with the associated moving of aged data between different storage tiers)

As mentioned in the previous section, the SAP HANA platform comprises memory-based and disk-based data stores. The platform supports relational data, text data, spatial data, series data, and graph-structured data and provides various processing options such as SQL, SQLScript, calculation views, and libraries for business calculations as well as predictive analysis. From the perspective of data flows, there are several typical patterns, some of which are explained in **Figure 9**.

01001010111010010
10100101000101010
01001010001011101
00101010010100010
10100100101000101

The integration services of the SAP HANA platform facilitate connection to virtually any data source and cover the requirements of lambda and other architecture principles that may be related to Big Data handling.

Figure 9: Typical Data Flows



vUDF = Virtual user-defined function
SQL = Structured Query Language
HDFS = Hadoop distributed file system

(A) Incoming streams of raw data can be filtered, analyzed, cleansed, and preaggregated in real time with smart data streaming in SAP HANA. Such preprocessing can, for example, be used for correcting outliers and missing values in streams of sensor data, or for detecting patterns that indicate alert conditions.

(B) Smart data streaming creates preprocessed and condensed high-value data, which is stored in in-memory tables (real-time views), for example, for real-time analytics and for combining with data created by enterprise applications on SAP HANA. Smart data streaming is capable of handling high volumes of rapidly incoming data of various types. Steps A and B correspond to the speed layer in the lambda architecture.

(C) In parallel, the incoming raw data is appended to a huge data store (C1) where all data is collected. Such massively growing data stores are often called data lakes. Data collected this way can be utilized for later analysis and information extraction, potentially using the complete data set, including historical data. HDFS is a scalable, fault-tolerant, and comparatively inexpensive data store, which makes it a suitable choice for storing this constantly growing set of incoming raw data. Stored in HDFS, the data can be further processed with a variety of data technologies from the Hadoop family. It can also be queried and analyzed with SAP HANA Vora. It should be mentioned, however, that raw data need not necessarily be stored in HDFS in all cases. If the raw data is structured, and if SQL is sufficient for processing, it can also be stored in the dynamic tiering extended store (C2).

(D) Smart data streaming is the preferred option in an SAP HANA platform environment. This is because it has some unique processing features and reduces the cost of development and operation with powerful development and administration tools, as well as integration into the SAP HANA platform. But if real-time processing of incoming data is not required, the raw data can be directly stored in HDFS. Data routing is managed through data collection technologies such as Kafka and Flume as alternatives to smart data streaming.

(E) The raw data is preprocessed on the Hadoop side by means of batch jobs, such as information extraction, transformations, filtering, data cleansing, preaggregation, or analysis of multimedia content. SAP HANA is also capable of executing the corresponding MapReduce jobs. But depending on the selected setup, this kind of batch processing (batch layer of the lambda architecture) can be executed on the Hadoop cluster outside the control of SAP HANA. This can be done using the various processing engines of the Hadoop family. The result of the preprocessing is again stored in HDFS and can be accessed from SAP HANA (see step F).

(F) Structured data can be directly read from HDFS files into SAP HANA. The virtual tables for smart data access can be used to execute federated queries against data in SAP HANA Vora, Spark SQL, or Hive. Virtual user-defined functions[6] (vUDFs) that trigger MapReduce jobs can be used if application-specific processing is required that cannot be done with SQL on Hadoop. Whenever the vUDF is called, the MapReduce job is executed on the Hadoop cluster and the result is returned to SAP HANA. Because this option is slower, results can be cached on Hadoop to improve the performance if the underlying data changes infrequently.

(G) High-value data created as the result of preprocessing within Hadoop (by batch, vUDF, or federated queries) can be stored in in-memory tables in SAP HANA for low latency access, for efficient processing with the database's native engines, and for combining it efficiently with other data in SAP HANA. Such derived high-value data can be pulled by SAP HANA (with vUDF calls and remote queries), or it can be pushed to SAP HANA from Hadoop – for example, over a JDBC connection. Storing the derived data in in-memory tables in SAP HANA has the advantage that the data can be accessed with very low latency; but depending on the use case, some mechanism for refreshing the data may need to be implemented. The alternative is to not persist the results in tables but to execute the remote queries or vUDFs each time the data is accessed by the application. The decision depends on how frequently the data is accessed, how fresh the data needs to be, and what access latency can be tolerated. Therefore, the option should be determined based on the specific use case.

(H) Data discovery tools can consume data from SAP HANA Vora through SAP HANA using smart data access (H1). It is also possible to connect such tools to SAP HANA Vora through Spark SQL for data discovery and visualization (H2), independently of the server for the SAP HANA database. That way, tools such as SAP Lumira® software can be used with SAP HANA Vora to analyze and visualize data in Hadoop. However, a decisive advantage of using SAP HANA is that its adapter for Spark SQL enables data analysts to combine data in SAP HANA Vora with business data in the SAP HANA database. With data virtualization through SAP HANA, any application or tool enabled for SAP HANA is automatically enabled for SAP HANA Vora, since the existing interfaces and connectivity are used.

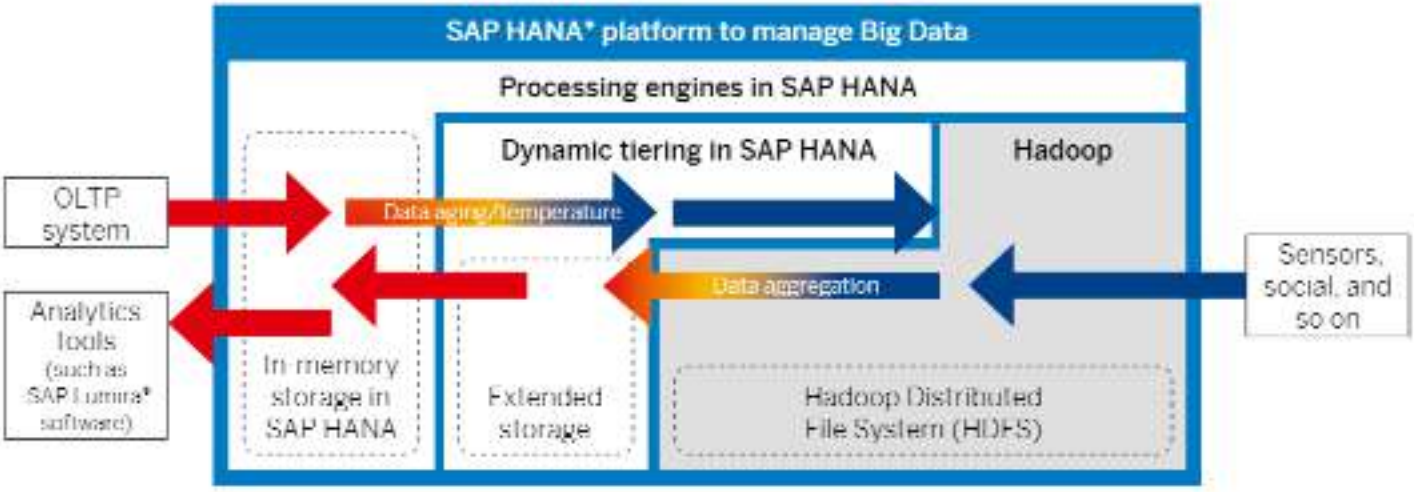
6. To deliver and run MapReduce jobs using SAP HANA, developers need to define a table valued virtual user defined function (vUDF). SAP HANA applications can use vUDF calls in the FROM clause of an SQL query such as a table or view. The vUDF is associated with a MapReduce Java program, which is shipped together with the function definition as SAP HANA database content.

(I) A business application on top of SAP HANA may also create data. Initially, the data is of high value and therefore stored in in-memory tables, where it can be accessed with very low latency. After some time, the data may become less relevant and is moved either to disk-based tables in the extended store (dynamic tiering) or to HDFS. In dynamic tiering the data can continue to be managed with tools in SAP HANA and queried with high performance, and selective updates are still supported. In contrast, data in HDFS can no longer be changed, but applications that run on the SAP HANA database can still query it through virtual tables when required.

(J) Data warehousing based on the SAP HANA platform, which includes the SAP Business Warehouse application and native modeling tools in SAP HANA, can easily consume Big Data through the platform, and can make the warehousing process simpler and more agile, virtual, and comprehensive. Particularly in landscapes consisting of many systems and data sources, a centralized data warehouse supports the combination of Big Data stored in the platform with centralized corporate data to achieve new insights. Data warehousing using SAP HANA creates a central place where one version of the truth is available, based on trusted, harmonized data.

As we have said, raw external data may come from a variety of sources – sensors and machines, social media content, e-mails, Web content, Web logs, security logs, text documents, multimedia files, and more. The data may be collected, analyzed, and aggregated in Hadoop, and the extracted high-value (hot) data is moved to in-memory tables in SAP HANA. This case is depicted in Figure 10 (from right to left) as the blue arrow that turns red.

Figure 10: Data Flows in Combined Data Processing Technologies



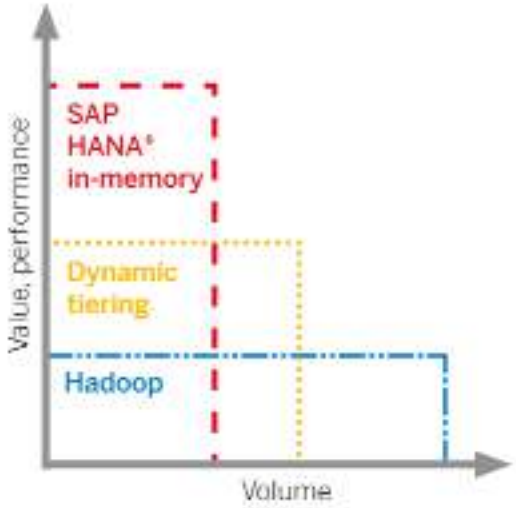
OLTP Online transaction processing

Figure 10 also shows that another typical data flow goes in the opposite direction (from left to right). In this scenario, the high-value (OLTP) data is created by an SAP business application and stored in in-memory tables in the SAP HANA database (red arrow), where it can be accessed with very low latency by analytics tools such as SAP Lumira. But not all data has the same business value, and not all data needs to be kept in-memory forever. Aside from its business value, data can be categorized in terms of volume, access patterns, and performance requirements. Based on all these characteristics, the different categories of data can be stored in different data stores with different storage capabilities, performance, and cost. When data loses value over time, it is said to get colder. It can then be moved to a different storage tier with higher latency, bigger capacity, and maybe less cost. In Figure 10, this is represented by the red arrow that turns blue.

When considering the dimensions value, volume, and processing performance for data, clear differences can be observed. In-memory storage in SAP HANA has the highest processing performance and is therefore used for high-value data, which is characterized by its comparatively low volume (up to several terabytes, though higher volumes can be reached). Higher volumes of warm and raw data are not relevant for in-memory storage but are best managed with dynamic tiering (for warm data) or Hadoop (for cold business data, raw data, or data of unknown value). Figure 11 depicts a qualitative comparison of these technologies in terms of data value, processing performance, and volume.

As we have seen, there are multiple options to deal with data of different types. SAP HANA is ready to handle data regardless of its characteristics in different storage tiers and with different processing engines – enabling it to act as the unifying data platform.

Figure 11: Qualitative Assessment of Data Processing Technologies



Big Data Use Cases

The data flow scenarios presented in the previous sections represent basic principles. Real-world Big Data use cases illustrate concrete areas of application. Possible new applications are emerging on a daily basis. Successful Big Data use cases are already being deployed in the following areas (to name just a few):

- Anticipating consumer behavior
- Increasing safety
- Mastering performance
- Redefining operational efficiency
- Managing predictive maintenance
- Preventing fraud
- Saving lives by improving medical research and services
- Personalizing real-time promotions
- Preventing injuries in sport
- Enhancing fan experience

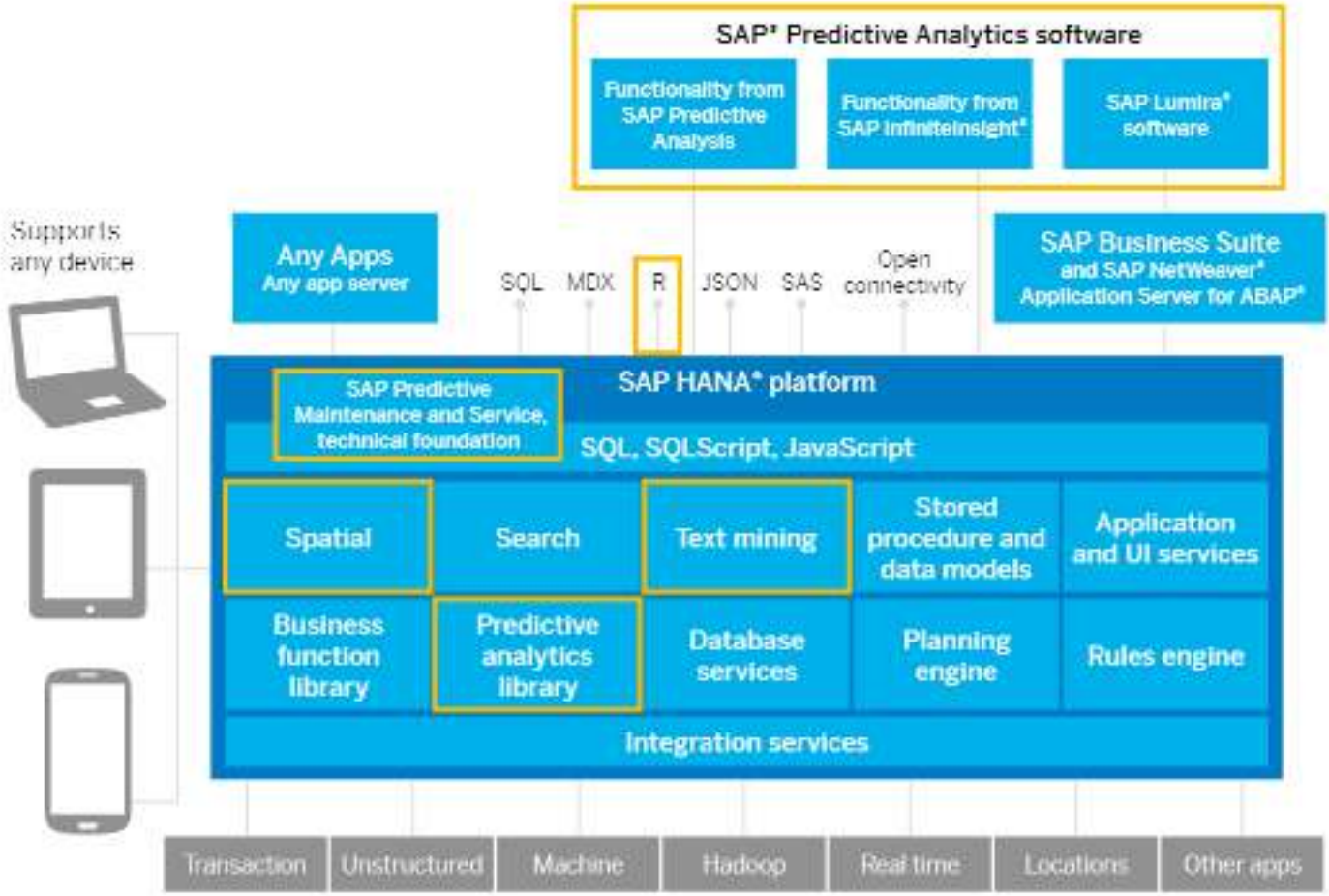
We will look at two examples that involve using SAP technology alone. However, since the SAP HANA platform is open, other components such as Hadoop and technologies from the Hadoop family can be incorporated into the setup as required. In any case, when designing the Big Data system landscape for your business, you should look for opportunities to simplify your IT infrastructure. And putting the SAP HANA platform at the center of your Big Data landscape is a strong start toward simplification.

PREDICTIVE MAINTENANCE AND SERVICE USE CASE

This use case is about obtaining meaningful views on data from machines, assets, and devices for making better real-time decisions and predictions and for improving operational performance. Typically, new business models from Industry 4.0 demand that companies improve their asset maintenance so as to achieve maximum machine availability with minimum costs. Companies also need to reduce their spare parts inventory, minimizing the amount of materials consumed by maintenance and repairs. This requires predictive analytics and algorithms to forecast equipment health.

The architecture setup in this case should allow real-time operations, analyses, and actions. At the same time, a very large number of events per day must be correlated with enterprise data. The SAP HANA platform, enhanced with the SAP Predictive Maintenance and Service solution, meets all the requirements associated with this use case (see Figure 12).

Figure 12: Predictive Maintenance and Service Setup



Note: Gold frames indicate particularly important components

SQL = Structured Query Language
MDX = Multidimensional eXpressions
JSON = JavaScript Object Notation
SAS = Statistical Analysis System

SAP Predictive Maintenance and Service enables equipment manufacturers and operators of machinery and assets to monitor machine health remotely, predict failures, and proactively maintain assets. It is offered as a standard cloud edition or as a repeatable custom solution for the technical foundation and a custom development project. Both options are based on SAP HANA.

Using a Lot of Data from Any System

To meet the requirements of this use case, your infrastructure must be able to use data from any system, SAP or non-SAP, regardless of the data type (business or sensor data). This is where SAP Data Services software is used. Furthermore, your infrastructure should be able to listen to 1 to 2 million records per second and only keep what is interesting. SAP Event Stream Processor handles this part. Through integration of Hadoop, the platform is able to connect to many terabytes, even petabytes, of back-end data.

Processing a Lot of Data

Predictive maintenance is an important step for keeping assets from failing. For that purpose, sensor, business, environmental, sentiment, and other data must be analyzed to discover relationships, patterns, rules, outliers, and root causes and thereby enable predictions. Based on this data mining, it is possible to take actions such as creating notifications, altering maintenance schedules, prepositioning spare parts, adjusting service scheduling, changing product specifications, and more. With its query engine and rules engine, the SAP HANA platform lets you perform ad hoc queries on a large number of records combined from many sources, in seconds. The SQL, SQLScript, and JavaScript capabilities allow utilization of fast programs that use many queries. With the built-in sophisticated algorithms of the predictive analytics library, you can carry out complex analyses. In addition, with the rules engine you can define and trigger business rules. You can also expand to planning scenarios by using the planning engine. Text analysis (search and relate) is no problem with the text analysis engine, which allows handling text as just another query aspect. Similarly, the geospatial analytics capabilities of the platform make comprehensive analyses of geospatial data possible. These engines enable you to find reasons for failure (root cause), detect deviations from the norm, and create a prediction model from sensor data and failures.

Open Connectivity

With its open connectivity, the SAP HANA Platform enables you to reuse your Statistical Analysis System (SAS) models with high-speed execution, should you need to connect to an SAS application. Moreover, you can explore statistics or even let the system make a best proposal by using SAP Predictive Analytics software. Because SAP HANA is able to deal with R procedures, if an R server is connected, for example, in order to use data mining algorithms for predictive stock and workforce optimization, you can reuse your R models. By integrating a Hadoop system, you can flexibly scale out to meet the requirements of preprocessing and storing vast amounts of data. Hadoop may also be used for archiving historical data or for offline batch processes.

This use case requires quick data correlations and actions to anticipate operational needs and unexpected breakdowns, and to automate triggers. In this system configuration, flexible predictive algorithms and tools combine technical and business data. Machine-to-machine communication helps monitor activities and stores data to fuel real-time reporting. Sophisticated business intelligence tools are required to obtain meaningful visualizations (such as SAP Lumira can provide).

DEMAND SIGNAL MANAGEMENT USE CASE

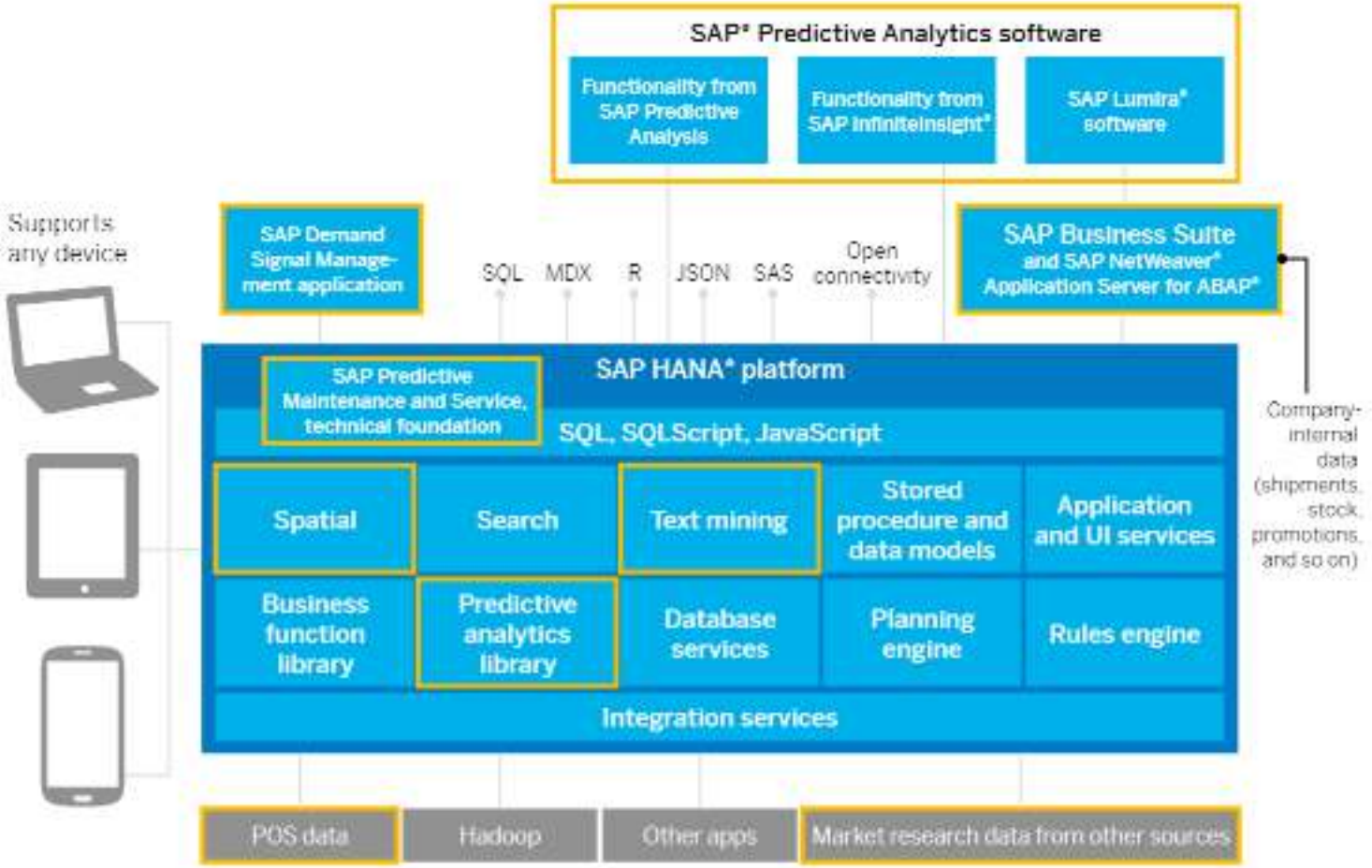
Demand signal management is a common requirement of consumer products companies. They are looking to apply their efforts toward those markets of retailers and consumers where they can realize the greatest growth. To realize this goal, they need a consistent and comprehensive global market view to be able to understand the demand from these various markets. Typically, these companies use syndicated data from various agencies to understand demand and brand perception so they can focus on the right areas. But if it is not automated, the consolidation and harmonization of data from various sources (internal and external) is a highly time-consuming and error-prone manual activity. The SAP Demand Signal Management application powered by SAP HANA addresses the most common challenges of data harmonization and automation for marketing, supply chain, and sales. With SAP Demand Signal Management, the SAP HANA platform can be evolved to include data sources such as weather data or social media data. SAP Demand Signal Management can act as a central platform for various use cases such as trade promotion optimization, sentiment analytics, demand forecast, and brand perception in consumer products and other industries. Figure 13 illustrates a typical demand signal management setup.

High-volume, high-variety, and high-velocity data can be processed both offline and in real time. The data can be stored in-memory or in the dynamic tiering option or Hadoop, depending on the requirements. Hadoop would be used, for example, when the volume of data coming particularly from external sources is extremely high and comes in with high velocity (as is the case with point-of-sales data).

In the retail industry, in the area of customer behavior analytics, a similar setup allows integration of several data sources into a central data store in SAP HANA, based on a predictive model that determines scores for several metrics such as return rate and likelihood to churn. The resulting scores can then be used in customer engagement intelligence for campaign target selection. The purpose is to improve the ROI of campaigns by targeting the right audience. In turn, this marketing optimization is expected not only to drive revenue and improve margins but also to deliver a personalized consumer experience across channels. A highly optimized analytical engine in SAP HANA enables processing huge amounts of data by using many different techniques to reduce the number of searches, as well as by employing the best scientific algorithms available.

In short, by deploying the SAP HANA platform, businesses are able to gain a distinct advantage over their competition. Powered by SAP HANA, SAP Demand Signal Management provides a centralized platform to monitor the aggregated market data, which in turn results in a better understanding of demand and the ability to focus in the right markets, thereby lowering costs and increasing revenues.

Figure 13: Demand Signal Management Setup



Note: Gold frames indicate particularly important components

SQL = Structured Query Language
MDX = Multidimensional eXpressions
JSON = JavaScript Object Notation
SAS = Statistical Analysis System
POS = Point of sale

SAP HANA Platform: Open, Flexible, Integrated, Scalable, and More

A UNIFYING DATA PLATFORM

SAP HANA is SAP's strategic platform for unifying and combining relational, text, spatial, series, and graph-structured data. It provides various processing options such as SQL, SQLScript, calculation views, and libraries for business calculations and predictive analysis. As described earlier, the platform comprises memory-based stores and a disk-based data store (the dynamic tiering option). To complete the picture, the SAP HANA platform comes with several options for integration with remote data sources, supporting data integration, data transformations, data quality operations, complex event processing, and federated query execution. SAP HANA puts a broad spectrum of capabilities at your disposal to manage the challenging characteristics of Big Data on one platform.

SAP HANA Vora also underlines the unifying character of the SAP HANA platform. This in-memory query processor has been built as an extension of Apache Spark and is effectively an in-memory query engine that can make the process of data analysis more in depth and oriented around business processes. Though it can be used independently of SAP HANA, businesses can benefit from using SAP HANA Vora as a bridge between regularly updated and accessed business data kept in SAP HANA with historical and mass data stored in Hadoop data lakes.

For all data, SAP HANA serves as the central point for data access, data modeling, and system administration. Having all data in one platform is a great advantage for application developers because they can access data in SAP HANA and external sources in a uniform way, such as with SAP SQL, SQLScript, and calculation views. The built-in rule engines and business functions accelerate application development. And having all data in a unified platform also facilitates analytics for real business value.

The SAP HANA platform helps your business meet the demands of the digital economy, where every company will be a technology company. A key challenge faced by all companies across every industry is driving innovation while tracking results in real time. Unlike traditional databases, SAP HANA is ready to manage this challenge successfully by removing data silos and providing a single platform for storing transaction, operational, warehousing, machine, event, and unstructured data – one real-time system operating on one copy of data, supporting any type of business workload, all at the same time.

OPEN AND FLEXIBLE

You can build a complete lambda architecture with components of the SAP HANA platform, including the processing engines and data stores. Further, because of its openness to integrating other technologies, SAP HANA lets you go beyond lambda and cover other reference architecture principles and requirements by combining different components. This makes the platform very flexible and therefore a sustainable investment.

SAP HANA has a powerful and flexible search function that allows searching across structured and unstructured data. Analysis of sentiment can be achieved by summarizing, classifying, and investigating text content. Persistence, manipulation, and analysis of network relationships and graphs can be carried out without data duplication.

INTEGRATED AND SCALABLE

Predictive analysis is enabled through prepackaged, state-of-the-art predictive algorithms that operate on current data, as well as through open predictive and machine-learning abilities, such as by integrating an R server. Interactive planning is possible without moving data to the application server. SAP HANA allows cleansing of locally generated and imported data, which can be performed without postprocessing.

Hadoop is suited for data that can grow infinitely – that is, for unstructured and raw data and when massive scale-out is required for processing. With Hadoop you can scale out flexibly. As a good complement, SAP HANA Vora is the recommended SQL engine for high-performance analytics on data in Hadoop and Spark. To achieve high performance, SAP HANA Vora uses advanced algorithms and data structures, and just-in-time compilation of query plans into machine-executable binary code. The integration of SAP HANA Vora into the Spark execution framework has the advantage of reusing various Spark capabilities, such as the Spark API, and Spark's ability to integrate with a cluster manager such as Hadoop YARN.[7] This integration makes it possible for Spark programs and Spark SQL queries to access data using SAP HANA Vora and combine it with other Spark data sources and processing modules. As mentioned earlier, while SAP HANA Vora can be used independently of the SAP HANA platform, it is optimally integrated in the SAP HANA platform to help ensure very high efficiency.

SIMPLIFICATION AND SECURITY

Using SAP HANA as the unifying platform for all your data also simplifies system administration and software lifecycle management, thus helping to reduce cost of ownership. You can gain efficiency and agility while reducing IT expenses by running SAP HANA in a virtualized environment. And the enterprise-class high-availability and disaster-recovery features in SAP HANA are designed for continuous operation, even if failures occur.

SAP HANA provides versatile tools such as the SAP HANA studio, SAP HANA cockpit, SAP DB Control Center systems console, SAP Solution Manager, and SAP Landscape Virtualization Management software to monitor the health of your system and effectively administer your data infrastructure. You can also manage the platform lifecycle more efficiently by streamlining installations, configurations, and upgrades for your entire SAP HANA platform environment, using another rich set of tools to help simplify deployment and maintenance while reducing costs.

7. YARN – Yet Another Resource Negotiator.

With SAP HANA you can set up simplified data warehousing that reduces IT workloads, enhances data modeling, and simplifies administration – particularly important for landscapes consisting of many systems and data sources. The centralized data warehouse not only consumes Big Data stored in the SAP HANA platform but also harmonizes it and combines it with centralized corporate data, thereby providing one version of the truth based on trusted data.

Last but not least, SAP HANA helps you ensure your business data is safe by providing security functions to implement specific security policies. Additionally, when you run your SAP applications on SAP HANA, they benefit from the same security foundation, and you can even add incremental protection. For example, you can control database administrators' access and integrate SAP HANA with your existing security infrastructure to help ensure your data is always secure.

CORE MANAGEMENT SYSTEM FOR BIG DATA

With SAP HANA, you benefit from intuitive, state-of-the art tools and technologies to effectively run a mission-critical, secure environment for your most valuable data assets. At the same time, you are preparing your business to manage the challenges of the digital economy.

For all these reasons, we recommend using SAP HANA as your core data management system for all Big Data applications, including custom applications, and adding other options such as SAP HANA Vora or Hadoop when additional capabilities are required.

SAP HANA enables you to implement a complete lambda architecture from preintegrated components of the platform, including the processing engines and data stores.

Find Out More

Here are sources for more information on the ways SAP can help you manage and get the most from Big Data:

- SAP HANA platform: **http://go.sap.com/solution/in-memory-platform.html**
- Administration and IT operations for SAP HANA: **http://hana.sap.com/capabilities/admin-ops.html**
- SAP HANA and Apache Hadoop: **www.sap.com/solution/big-data/software/hadoop/index.html**
- SAP HANA Vora: **http://go.sap.com/product/data-mgmt/hana-vora-hadoop.html**
- SAP IQ for logical Big Data warehousing (OLAP): **www.sap.com/iq**
- SAP SQL Anywhere for designing embedded database applications for mobile and remote environments: **www.sap.com/pc/tech/database/software/sybase-sql-anywhere/index.html**
- SAP Data Services for all types of data integration, including smart data access, smart data integration, and smart data quality: **www.sap.com/pc/tech/enterprise-information-management/software/data-services/index.html**
- SAP HANA smart data streaming: **http://help.sap.com/hana_options_sds**
- SAP Event Stream Processor – a complex event processing platform for event-driven analytics: **www.sap.com/pc/tech/database/software/sybase-complex-event-processing/index.html**
- Advanced analytics with SAP HANA for in-memory processing for text, spatial, graph, and predictive analysis: **http://hana.sap.com/abouthana/hana-features/processing-capabilities.html**
- The Data Science organization at SAP, with experts who help implement predictive analytics and custom apps: **www.sap.com/solution/big-data/software/data-science/index.html**
- SAP Predictive Maintenance and Service, cloud edition: **http://help.sap.com/pdm-od**

Learn more about SAP HANA for the Mission

CONTACT

www.sapns2.com

info@sapns2.com

877.972.7672

NS2®

SAP NATIONAL SECURITY SERVICES

Now, let us help you. Learn more at sapns2.com/