Executive Summary

SAP HANA is already helping businesses extract value out of Big Data by enabling real-time analysis of diverse data sets. Apache Hadoop is being used today to manage petabytes and possibly exabytes of data. SAP HANA uses in-memory computing to deliver analytic results much faster than Apache Hadoop. But Hadoop can handle arbitrary data structures more easily and at much lower hardware storage costs per terabyte than traditional data warehouses. Clearly, SAP HANA and Hadoop have complementary strengths. So when you put SAP HANA and Hadoop together you get the ability to analyze big data in real-time, allowing you to make better decisions faster.

To help enterprises worldwide take full advantage of the combination of these technologies, SAP and Intel engineers worked together to integrate and optimize the Intel® Distribution for Apache Hadoop with SAP HANA. This document describes the reference architecture of the real-time big data analytics solution jointly engineered by SAP and Intel to deliver the scalability, performance, security, and manageability required by enterprises whose businesses depend on the quality and timeliness of data-driven decisions.

Contents

Starting with a Representative Use Case ................................................................. 2
Defining System Requirements ............................................................. 2
Combining SAP HANA with the Intel Distribution for Apache Hadoop .............. 3
Optimizing for Performance, Security, and Manageability .............................. 6
Conclusion .............................................................................................................. 7
Starting with a Representative Use Case

For over ten thousand years, agriculture has depended upon the farmer’s ability to analyze weather reports, soil conditions, terrain maps, water resources, and market conditions. Yet, it is only in the last twenty years that open access to accurate real-time data has unlocked dramatic improvements in modern agriculture. Tractors, tillers, combines and other agricultural machines now use embedded sensors to provide real-time data continuously. These smart machines can assess environmental conditions and communicate with each other. The real-time analysis of this cornucopia of data can help today’s farmer predict crop yields, increase the efficiency and reduce the cost of their equipment, develop better strategies for mitigating inevitable risks, and optimize yield prices.

Consider an agricultural machine manufacturer that wants to provide agronomics -- agricultural data-driven analytics -- to help farmers improve their business. The manufacturer has built sensors into the field equipment. By combining sensor data from the equipment with detailed data on field, weather, and market conditions, the company can create a rich and full picture of the farm and its environment. The company can then use advanced analytics to generate recommendations for each farmer. The benefits to the farmer are considerable. Since the recommendation is specific to time and location -- at the granularity of the area of the field and day of the week -- and can specify soil treatments – water, fertilizer, and pesticides – to improve crop yields, the farmer is now equipped with much more than an agricultural machine -- they now have an intelligent system backed by real-time big data analytics. Because the agronomic system fosters better use of the farm equipment and provides detailed and continuous knowledge of the customer, the machine manufacturer benefits by growing a loyal customer base and sustainable communities that can afford newer equipment.

Defining System Requirements

The business requirements of the agricultural machine manufacturer are as follows:

- **Capture** data from various data sources: equipment sensors, soil conditions, crop history, plant data, weather data, climate history, market data, and other relevant sources.
- **Store** the large and growing volume of data in a scalable data reservoir
- **Cleanse** and prepare data for analysis
- **Analyze** the data by executing a variety of statistical methods. The analytics can be further subdivided:
  - Identify the problem (formulate the query)
  - Execute the query
  - Draw conclusions
  - Revise the query
- **Present** the data to the farmer through intuitive analytic interfaces or BI tools
The technical requirements of the system would stretch the abilities of traditional data warehouses and conventional RDBMS but the advent of In-memory database computing and distributed data management technologies offers the critical building blocks:

- **Data Stores**
  - Distributed Data Store for large volumes of unstructured data (sensor logs, history)
  - In-memory Database for streaming data (weather data, market data)
- **Fast Analytics** (Real-time response to queries on some portion of the data)
- **Deep Analytics** (Advanced analytics across entire data set)
- **High-performance Connector between the Analytics engines**

Combining SAP HANA with the Intel Distribution for Apache Hadoop

Understanding the differences in capability between SAP HANA and the Intel Distribution for Apache Hadoop is the first step to recognizing the value in their integration. SAP HANA is already helping businesses extract value out of Big Data by enabling real-time analysis of diverse data sets. Likewise, Apache Hadoop is being used to manage petabytes and quite possibly exabytes of data, far more than the 100 TB or less that SAP HANA and traditional relational database management systems (RDBMS) typically handle. Hadoop is significantly slower than SAP HANA, taking minutes or hours to provide analytic results, but it can handle arbitrary data structures more easily and at much lower hardware storage costs per terabyte than traditional data warehouses. SAP HANA and Hadoop have complementary strengths. So when you put SAP HANA and Hadoop together you get the ability to analyze big data in real-time, helping you make even better decisions faster.

<table>
<thead>
<tr>
<th>SAP HANA</th>
<th>Intel Distribution for Apache Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainly structured data in memory</td>
<td>Any data or file structure on disk</td>
</tr>
<tr>
<td>Very fast access (~&lt;1 ms)</td>
<td>Very slow data access (seconds to hours)</td>
</tr>
<tr>
<td>Predefined schema</td>
<td>No schema or schema on read</td>
</tr>
<tr>
<td>1 or many servers (100s of cores)</td>
<td>Cluster of servers (1000s of cores)</td>
</tr>
<tr>
<td>Scale-up or scale-out architecture</td>
<td>Scale-out architecture</td>
</tr>
</tbody>
</table>
The use of Hadoop does not fundamentally change the type of analytics that can be performed with SAP HANA. The use of Hadoop does, however, change the scope and architecture of the analytics solution. Because the data volumes stored by Hadoop can be so vast, its raw data cannot be readily copied into an existing disk-based or in-memory data warehouse. It is too much data to economically store permanently, and moving data from Hadoop just in time for analysis would take too long. This means that some of the data analysis must be carried out in Hadoop.

Moreover, queries executed in Hadoop usually take longer to run, anywhere from minutes to hours. The solution must compensate for the latency or else the analytic process of “formulate query -> analyze data -> present results -> draw conclusions” will slow to the point where it cannot be done in real time. And finally, some of the analytic operations will run on multiple analytic engines that each draws from a different data store, consolidating the results from the in-memory database, the on-disk data warehouse, and the Hadoop cluster. This usage is the focus of the integration between the Intel Distribution for Apache Hadoop with SAP HANA.

**Federating Queries**

A federated query is one in which the original query is split into separate parts that can run on any combination of the Intel Distribution for Apache Hadoop, SAP HANA, SAP Sybase IQ, and other data platforms. The parts are then executed separately.

A query running on the Intel Distribution, for example, may have to perform a database “join” with data in SAP HANA. This presents two options: retrieve data from SAP HANA while the Hadoop query is running or copy the data in SAP HANA to Hadoop before the analysis query on Hadoop starts. There may also be dependencies between the parts of the query. For example, a job running on the
Intel Distribution may need to complete first so that the results it produces are available to complete a query in SAP HANA. Because one part can take a longer time to run, the federated query must have a way of executing queries asynchronously, where the system accepts the query and passes control back to the user so they can continue with other activities. Later, when the query is complete, the system notifies the user so they can examine the result at that time.

SAP HANA integrated with the Intel Distribution for Apache Hadoop implements federated queries in four ways:

1. **Client-side Federation** (available in BI 4.0 FP3): SAP Business Objects BI executes queries separately and consolidates data from the Intel Distribution and the SAP HANA database.
2. **Standalone Analytics**: SAP Business Objects BI analyzes data in Hadoop directly.
3. **Query Federation** (available in SAP HANA SP06 with SAP HANA smart data access): SAP HANA executes queries on multiple data sources including the Intel Distribution for Apache Hadoop, consolidating and combining the analytic results.
4. **Data Federation** (available with SAP Data Services 4.1): SAP HANA treats data in the Intel Distribution as an external table.
Optimizing Query Federation for Performance and Security

In SAP HANA integrated with the Intel Distribution for Apache Hadoop, the optimization is dual-fold: Intel software developers optimized both SAP HANA and the Apache Hadoop platform for optimum performance on Intel Xeon processor-based servers. SAP HANA takes full advantage of advanced features in the Intel® Xeon® processor E7 family, such as expanded memory addressability, increased core count, larger cache, as well as security, reliability, availability, and serviceability features. The Intel® Distribution for Apache Hadoop* software is built from the silicon up to deliver security with no compromise on performance and manageability by making full use of the Intel Xeon processor, SSD storage, and 10GbE networking.

The collaboration between SAP and Intel has been even more fruitful in the joint engineering of the SAP Smart Data Access Connectors optimized with the Intel Distribution for Apache Hadoop. The connectors allow the Intel Distribution to transcend the traditional use of Apache Hadoop as an engine for offline batch processing.

In SAP HANA SP06, the Smart Data Access connector enables the Intel Distribution for Apache Hadoop to operate as an extended data store for SAP HANA using the proxy table construct. Intel and SAP jointly engineered the connector with several unique optimizations such as caching and hot data replication to improve performance. In addition, the connector uniquely implements end-to-end security with encryption of data in motion and at rest. When a user submits a query via SAP HANA studio, the combined solution returns the result transparently with no loss of real-time performance.

In SAP HANA SP07, the integration is further enhanced to support data relocation between SAP HANA and the Intel Distribution. When a query -- for example, to join a smaller data set in SAP HANA with a larger data set in the Intel Distribution for Apache Hadoop -- is submitted to SAP HANA, the connector relocates the data to the Hadoop cluster, where the job is executed. The Intel Distribution then returns the result to SAP HANA via the proxy table.

SAP HANA integrated with the Intel Distribution for Apache Hadoop provides a real-time big data analytics solution. Unlike other offerings in the industry, the SAP HANA solution with the Intel Distribution is optimized top-to-bottom through the stack with Intel technologies such as data security enhanced by Intel AES-NI and Intel Expressway Gateway, Intel Manager equipped with Intel Active Tuner, and balanced performance across Intel SSD storage and 10GbE networking.
The Intel® Distribution for Apache Hadoop software (Intel® Distribution) includes Apache Hadoop along with other software components enhanced and optimized by Intel for performance, security, and manageability.

- **Performance:** The Intel Distribution 2.5.1 has been shown to perform 2.4X better than the HortonWorks Data Platform 1.3.9 on the same hardware, completing the Terasort and Wordcount workloads on a 10-node cluster in significantly less time.

- **Security:** Only the Intel Distribution is built from silicon up to deliver security without compromising performance and manageability. The Intel Distribution running on Intel® Xeon processors enables Hadoop to make full use of the performance and security capabilities available in the processor. For example, the Intel Distribution uses Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI) available in Intel Xeon processors to accelerate cryptographic functions, erasing the typical performance penalty of encryption and decryption. Using AES-NI, the Intel Distribution has been shown to deliver up to 20X increases in throughput over the same operations performed without the use of AES-NI.

- **Manageability:** Only the Intel Distribution has an automated performance tuning mechanism that saves configuration time while selecting a more optimal configuration than a human expert. The Intel Active Tuner uses a patented mechanism to test various parameters rapidly until it converges on the optimal configuration for a given Hadoop application.

The Intel Distribution is developed and supported by a software team that has real-world experience with Hadoop deployments as well as deep expertise in the full platform stack of Apache Hadoop on Intel hardware including Java and Linux. Moreover, Intel contributes its knowledge and all hardware-enabled platform enhancements to the open source Apache Hadoop community.

### Conclusion

SAP is uniquely positioned to address the needs of the enterprise business analysts with a platform that can deliver real-time analytics operating on a diverse range of data sources. SAP HANA, by virtue of its in-memory database computing engine, already supports a variety of business applications that require instant analysis of data stored in transactional databases. Apache Hadoop complements and enhances this state-of-art by enabling access to a potentially unlimited source of unstructured data.

By integrating SAP HANA with Apache Hadoop, SAP can provide customers -- such as the agricultural machine manufacturer -- with a real-time big data analytics solution that scales with the volume of data. By selecting the Intel Distribution for Apache Hadoop as its preferred platform, SAP relies on Intel to deliver performance and security enhanced by Intel hardware, and to provide an open source data management platform on which SAP can innovate. By signing an agreement to resell the Intel Distribution, SAP provides its customers with a complete set of big data software and a single point of contact for technical support.
For more information, visit [http://hadoop.intel.com](http://hadoop.intel.com).

To get started with a proof of concept or pilot project with SAP HANA and the Intel Distribution for Apache Hadoop, contact SAP sales or Intel Software Sales.