Blueshift RESEARCH **INITIAL REPORT**

Guido Gualandi, *gg@blueshiftideas.com*
Reverdy Johnson, *rj@blueshiftideas.com*, 415.364.3782

Big Data Market Fragmented, SAP's HANA Praised

Companies: EMC, EPA:DSY, HPQ, IBM, INFA, MSFT, MSTR, ORCL, QLIK, SAP, SOW, TDC, TIBX **April 3, 2012**

Research Question:

Which companies have the best tools to connect and manage big data from the Web to SAP's HANA and other analytics platforms?

Summary of Findings

- Ten of 14 primary sources said the market for big data management and analysis tools is fragmented as many solutions exist and companies' data and storage needs vary greatly. The nascent big data market may need time to mature before a dominant player emerges.
- Most SAP AG (SAP) and Oracle Corp. (ORCL) clients will utilize these companies' tools for a streamlined approach when they need to work primarily within their systems.
- Six sources said Informatica Corp. (INFA) is the leader among companies working in non-SAP or non-Oracle environments in obtaining data, cleaning it and making it available to any analytics software. Sources touted Informatica's complete solution, ease of use and cost-effectiveness.
- IBM Corp.'s (IBM) Netezza was quoted as one of the best solutions for working with unstructured data.
- Nine sources praised SAP's new HANA data platform. It is gaining positive word of mouth from proofs of concept and stands to take share from Oracle, especially among customers using both Oracle and SAP products.
- Hadoop's open-source software is an up-and-coming solution that allows companies to retrieve and work with unstructured data. It can be used with open-source ETL tools or together with Informatica.

	Competition Rising, Market Fragmented	INFA in Non-SAP/Oracle Environments	SAP's HANA
IT Department Personnel	↑	↑	↑
IT Consultants	↑	↑	↑
Industry Specialists	↑	↑	↑
Database Consultants	↑	↑	↑

Silo Summaries

1) IT DEPARTMENT PERSONNEL

These three sources said big-data users are faced with the dilemma of staying with one vendor for every database solution and tool, or diversifying to choose best-of-breed in each category. **One source uses and praises Informatica as an independent solution and market leader. Two sources use Oracle databases but are considering switching to in-memory solutions; one is strongly leaning toward SAP's HANA.** Using server farms in a cloud with open-source software such as Hadoop is cheaper from a TCO point of view but presents customization and maintenance complications.

2) IT CONSULTANTS

Companies tend to use connectors and data tools from their main vendors, especially Oracle and SAP. **Informatica shines among companies not tethered to Oracle and SAP. Three of four sources praised Informatica as having the best tools for specific needs in a heterogeneous environment.** MicroStrategy is strong when partnered with Teradata and Informatica among companies that are not 100% SAP or Oracle. SAS, Talend and Hadoop also offer good tools, especially in the midmarket. Two sources said HANA is the most powerful database platform that will only get stronger and more prolific once it adds new applications in the coming year.

3) INDUSTRY SPECIALISTS

The market is fragmented based on the various needs and uses of analytical tools for managing big data, giving rise to many potential players and no overwhelming favorite. Social listening and Web data mining are done mainly in the Hadoop environment. Some of this data is cleaned and sorted with data quality management (DQM) software such as **Informatica, which was noted as a leader in this space because of its complete solution**. Data storage leaders include Oracle, IBM and Teradata. Teradata offers a good tool to work with large amount of data while MicroStrategy can act as the BI tool to handle large volume.

4) DATABASE CONSULTANTS

The field is crowded with quality offerings from IBM, SAP, Oracle, HP, EMC and Teradata. **Teradata together with MicroStrategy and Informatica are best-of-breed and liked because they are independent. SAP is gaining market share because its clients with an Oracle database are considering a database change in light of real-time computing and in-memory solutions.** The midmarket has no clear leader.

Background

Data growth is exploding and is being driven by company processes, social media and the Internet. In order for companies to react faster to the rapidly changing market environment, they need both the ability to amass the data and the processing power to analyze it. The usage of specialized hardware and software to store and process these massive amounts of data (aka "big data") is increasing. SAP and Oracle have new database machines designed specifically for this purpose. Blueshift's March 1 report on SAP's HANA found that it was performing well in pilot programs, is one to two years away from full implementation, and poses a threat to Oracle.

CURRENT RESEARCH

This report aims to uncover which companies have the best tools to address big data needs and to further understand these tools' growth potential. Blueshift employed its pattern mining approach to establish and interview sources in five independent silos:

1) IT department personnel (3)
2) IT consultants (4)
3) Industry specialists (4)
4) Database consultants (3)
5) Secondary sources (8)

We interviewed 14 primary sources and included eight of the most relevant secondary sources focused on the broader big data market, growth of Hadoop, analytical processing systems, two wins for Informatica, an SAP application allowing access to HANA from Apple Inc.'s (AAPL) iPad, and Teradata's challenge to Oracle's Exalytics.

Silos

1) IT DEPARTMENT PERSONNEL

These three sources said big-data users are faced with the dilemma of staying with one vendor for every database solution and tool, or diversifying to choose best-of-breed in each category. Going with one vendor, usually larger ones like SAP or Oracle, is tempting because doing so can reduce problems by maintaining one point of contact if anything goes wrong. However, not all large companies have the software needed in each field, and their solutions can be expensive. One source uses and praises Informatica as an independent solution and market leader. An SAP client expects to remain with SAP tools, while an Oracle client is leaning toward staying with Oracle tools but understands the advantages of switching to new tools. Two sources use Oracle databases but are considering switching to in-memory solutions; one is strongly leaning toward SAP's HANA. Using server farms in a cloud with open-source software such as Hadoop is cheaper from a TCO point of view but presents customization and maintenance complications.

➤ **Fabrice Benaut, CIO at IFR Monitoring, a GfK group specializing in marketing research on technical consumer goods**

This source uses Informatica and believes it is the market leader in providing tools to analyze big data. The system is easy to learn and requires only a small team. Informatica also has the advantage of being independent and versatile enough to work on all databases. IBM's Netezza is another quality solution, as is Qlik Technologies Inc.'s (QLIK) QlikView and SAP's Business Objects. Talend is the best of the open source options, but this source was hesitant to use open source tools because they require more customization and are too "open" for large enterprises such as his. He is considering using HANA as a database machine and working with data in the cloud as well.

- "We chose Informatica PowerCenter and Informatica B2B and data quality tools as they are the leader in the market. Their solutions made it possible to treat automatically large and complex data."

> Informatica is for sure No. 1 in working with big data and for their data quality tools, master data management tools and data exchange. They are expensive, but they are worth the price and you can also negotiate with them.
>
> *CIO, IFR Monitoring*

- "With Informatica, we were able to quickly discover and analyze data using prebuilt rules and a single development environment, and to reuse data profiling results across projects."
- "Using prebuilt Informatica tools is useful as you can have a small team and you can learn the tools quite quickly. You can also do complex operations if you need to. We were also able to work on complex information such as long texts."
- "Informatica is an independent company and connects with any software on the market in both directions. You can use the data in analytics platforms such as Netezza, QlikView or SAP's Business Objects [BO or BOBJ]."
- "There are two categories of ETL [extract, transform, load] software companies: dedicated ones like Informatica and IBM, which is a close second, and the low-end ETL solutions such as Microsoft [Corp./MSFT] or SAP's BO. Informatica is for sure No. 1 in working with big data and for their data quality tools, master data management [MDM] tools and data exchange. They are expensive, but they are worth the price and you can also negotiate with them."
- "Concerning open source, Talend is the best option, but it does not come free as you have to buy services and do customization. When you have Talend you also need to work on the solution, and in the end you are prisoner of the guy who did the development. It is difficult to use open source in large companies as we need solutions that are durable and not dependent on some IT person or developers. Open source is so open anybody can transform the solution, and that can be dangerous for a company."
- "IBM Netezza is a very good solution. It is most powerful in treating useful data, using streaming technology, and the best combination of hardware and software to get the highest performance. It can be a competitor of MicroStrategy [Inc./MSTR] or HANA and can be a good pair with Informatica tools."
- "EMC [Corp./EMC] and HP [Hewlett-Packard Co./HPQ] arrived later in this market and are a step behind. We also like QlikView as an in-memory Business Intelligence [BI] solution. It works a little bit like Cognos cubes a while ago but with more efficiency. The best feature is that it is not hierarchical, and you can navigate data in the way you want. Now they have consolidated their support team and improved their structure so that they can assist their clients."
- "We are going to look next at SAP HANA, but we don't have an opinion yet. We are Business Object clients."
- "Another area we are going to look at is analytic search such as [Dassault Systemes S.A.'s/EPA:DSY] Exalead and virtual MDM. It would be useful to work data in the cloud as we cannot store everything we find, but we need to find parameters so our questions have the same answers when asked in different moments."

> It is difficult to use open source in large companies as we need solutions that are durable and not dependent on some IT person or developers. Open source is so open anybody can transform the solution, and that can be dangerous for a company.
>
> *CIO, IFR Monitoring*

Business information specialist for an aerospace and defense company

This source's company is likely to use SAP tools for big data when the need becomes more urgent. It already uses SAP in other areas. Informatica and SAS have quality options but staying with SAP offers a streamlined solution. The company is considering changing from its Oracle database to SAP's HANA, which it deemed as superior and speedier.

- "We have not yet decided on a strategy for big data as it is not so urgent. Since we work with SAP, we usually wait for SAP to have the right tool as they will develop it sooner or later."
- "We have tried HANA, and it is an expensive but fast tool. Because we have mainly SAP tools, for us it makes sense to use SAP connectors to HANA or to our BI tools. BOBJ data services do the job of working with structured data."
- "Informatica and SAS [Institute Inc.] have very strong tools, but at the moment we are not looking to diversify. We have another component and we use other vendors in other subsidiaries, but we would rather migrate all of them to SAP tools, when we have the budget."
- "One big issue is about keeping or changing our Oracle database in favor of SAP HANA. Certainly, a lot of work needs to be done at database level if we want to increase speed."
- "Technologically HANA is superior to all other software we tried, and I think that is the direction we are going toward. Internal demand for

> One big issue is about keeping or changing our Oracle database in favor of SAP HANA. Certainly, a lot of work needs to be done at database level if we want to increase speed.
>
> *Business Information Specialist*
> *Aerospace & Defense Company*

real-time analysis comes mainly from the sales department. If we do something, it will be in that area. It will most likely be an all-SAP solution, but we might use some small applications to integrate from the Web if needed."

- IT manager for a large retail company in the EMEA
  As an Oracle client, this source is likely to stay with Oracle tools for big data. Still, it is considering other options, including a cloud solution or Hadoop, because Oracle's Exadata is expensive. Diversification comes with its own expenses because of the need for new or additional resources to manage the new tools. The source's company is waiting to make a decision and strongly considering an in-memory solution for its database bottlenecks.
  - "We are Oracle clients, and we mostly have Oracle tools. Oracle has good connectors also for big data. Our main problem is some bottlenecks at the database level. We think we can solve the problem with Hyperion and all Oracle middleware."
  - "However, we are hesitant about the solution we could use. On one hand, we need to work on the database and purchase Exadata and maybe Exalogic, but it is expensive. On the other hand, we could save money by having a different database or even Oracle database running in a cloud, maybe using Hadoop."
  - "It is quite confusing as there are many tools available. In the high end it would be easier with EMC, IBM and Teradata [Corp./TDC], but if you want a cheaper solution, you need a lot of knowledge about those relatively new Hadoop tools."
  - "The question of going with one vendor, Oracle in our case, is a good one concerning IT management. Diversifying is tempting, but we don't have resources in-house to manage new tools. We are afraid it could be much more expensive. Right now we are still in a wait-and-see mode, but we definitely want to go with an in-memory/real-time solution to get rid of our main bottlenecks."

> It is quite confusing as there are many tools available. In the high end it would be easier with EMC, IBM and Teradata, but if you want a cheaper solution, you need a lot of knowledge about those relatively new Hadoop tools.
>
> *IT Manager*
> *Large Retail Company in the EMEA*

2) IT CONSULTANTS

Companies tend to use connectors and data tools from their main vendors, especially Oracle and SAP. All companies need to retrieve and clean data, which is where Informatica shines among companies not tethered to Oracle and SAP. Three of four sources praised Informatica as having the best tools for specific needs in a heterogeneous environment. The market is becoming increasingly crowded and fragmented while companies' needs and various providers' solutions have become more specific. This is causing some companies to take longer with the evaluation process, allowing new market entrants time to develop solutions. MicroStrategy is strong when partnered with Teradata and Informatica among companies that are not 100% SAP or Oracle. SAS, Talend and Hadoop also offer good tools, especially in the midmarket. One source said Hadoop is the solution of choice because of its lower cost, greater scalability and more advanced open source solution. Two sources said HANA is the most powerful database platform that will only get stronger and more prolific once it adds new applications in the coming year.

- BI implementation specialist with an EMEA consultant
  A company that is dedicated to SAP or Oracle will choose the corresponding tools to maintain consistency, and may even wait for a tool to be developed rather than use an outside entity. IBM, Informatica and SAS all have quality alternative offerings, and Informatica offers superior tools for specific areas; this may appeal to an independent company but likely not an all-SAP or Oracle company. HANA is the most powerful reporting and analytics database machine and will continue to gain a following once additional applications are released early in 2013, putting SAP in position to lead the overall market.
  - "The choice of a solution usually follows what kind of software is mostly used in a company. For example, a company that is 100% SAP will choose to stay with SAP tools. Companies that are 80% with SAP or Oracle usually go with the vendor that is most common. It is not unusual to find companies who decide to consolidate on one vendor, and when they do it, even if it is painful, they do it with the vendor who is the most used. Some companies even decide to wait for the vendor of choice to develop the right tool if they don't have it available

yet. But if clients have platforms different from SAP or Oracle they will use the best-of-breed solutions such as Informatica."

- "For SAP clients, I usually suggest to use SAP ETL tools. HANA works well with BOBJ data services. If you have unstructured data, you can really do everything with SAP. There are data quality tools, data integration and replication tools—all you need, really. So it does not make sense to look elsewhere. However, when we have an all-Oracle client, we usually suggest to stay with Oracle."
- "Some large companies have already purchased other solutions, and in that case we also see IBM, SAS and Informatica. It depends on the cost analysis/TCO/ROI and what makes sense. For an SAP client, SAP solutions will have the best TCO and ROI, but for a client who has a variety of software solutions, Informatica and others make more sense."
- "HANA is the most powerful tool I have seen so far for reporting and analytics. As soon as SAP releases more software for HANA, which we expect at the beginning of next year, it will be No. 1 for rapidity in all fields. We have done some work with Oracle and Hyperion as well. That is good too but technologically inferior to HANA. For some clients with little data volume, QlikView can also be useful."
- "Similar to what we do with analytics, we usually stay with what the client uses most. Informatica has one of the best tools, but when you are 100% SAP you should stay with SAP MDM. Informatica is only a threat in multi-environment clients and, for us, that is not the majority."

> A company that is 100% SAP will choose to stay with SAP tools. Companies that are 80% with SAP or Oracle usually go with the vendor that is most common. … Some companies even decide to wait for the vendor of choice to develop the right tool if they don't have it available yet. But if clients have platforms different from SAP or Oracle they will use the best-of-breed solutions such as Informatica.
>
> *BI Implementation Specialist*
> *EMEA Consultantcy*

- "Oracle, IBM and SAP all have everything you need and, being big, they can cover pretty much everything. Smaller and independent companies such as Informatica have better tools in specific fields and can be used in certain cases like non-SAP environments or where SAP is less than 70%."
- "SAP will win in all segments. They have HANA, which is the most powerful tool, and when applications for HANA are released, companies will start migrating to HANA. At that point most clients will use SAP middleware and BI tools to work with HANA as consolidating on one vendor is the most efficient way. I believe that SAP is the best positioned of all vendors because they have the largest installed base and the most powerful tool."
- "Many vendors benefit with the increase in big data management. The large ones, Oracle, IBM and SAP, benefit because of their large installed base. Independents benefit because they offer solutions that are unique and solve specific issues the big ones can't. SAP is the leader now with HANA; they have the most potential. Informatica and SAS have valid tools as well and will continue to sell. However, I do not see SAP clients not using SAP tools and, therefore, SAP will continue to be the leader because they have the largest base."

> I believe that SAP is the best positioned of all vendors because they have the largest installed base and the most powerful tool.
>
> *BI Implementation Specialist*
> *EMEA Consultantcy*

➤ **BI specialist with an EMEA integrator, leading all implementations of BI software and connectors**
Informatica tools are frequently used, and its superior MDM platform is best for clients with an independent environment. SAP and Oracle still are the most frequently adopted solutions among their own clients. The source reported tremendous competition and numerous quality offerings from all major players, including SAS, HP, MicroStrategy, IBM, Qlik and Talend. The market is somewhat fragmented because of all the available choices and specific needs. Platforms that work well with Hadoop are becoming an important consideration as well. HANA will be the most dominant database machine because of its in-memory functionality.

- "Our clients use a variety of tools. Informatica PowerCenter or IBM InfoSphere DataStage are common, but we see SAP, Oracle, Talend, SAS and others. You need tools to clean data, verify their quality … and send them to the right application. You then can analyze the data with SAP BOBJ, IBM Cognos or MicroStrategy. HANA is just an in-memory database, and it does not do that job."

- "There are so many different possibilities. For example, if you have a big volume of data, Teradata appliances with MicroStrategy BI tools work really well. Teradata is excellent if you have work with petabytes and have complex calculations. HANA is really fast with volume as well."
- "If you want to analyze blogs, you will need some intelligent search with HP Autonomy or Exalead; there you can index all the information and send it to a BI tool to do a report."
- "QlikView and Tableau [Software] are quite good in a small company with limited data to analyze, but they will not replace business warehouse software. In the future, HANA will kill them both as it is faster and has broader functionalities while working in-memory as well."
- "Oracle Exalytics is expensive and relatively unknown. I haven't seen any in use yet. Informatica is also used to work with data from different environments. IBM Netezza is also an excellent tool, mostly used by IBM clients. In the end, SAP will win in real-time analytics as they have the best technology with HANA and the largest installed base in ERP."
- "Microsoft SQL Server 2012 integrated with SmartPoint [Technologies Ltd.] is adopted by many companies with Excel for reporting. It is cheap but works for many. The Exadata and Teradata platforms for large quantity of data are both good and expensive."
- "I am looking closely at all platforms that work well with Hadoop Map Reduce open-source software. IBM DB2-based Smart Analytics System and Netezza offerings and HP Vertica/Autonomy are strong players in this field. These are just some of the options on top of the usual SAP and Oracle."
- "For data analysis we still have to look at the BI players like SAP BOBJ, Oracle Hyperion or MicroStrategy. For small volume, we also see QlikView and Tableau. Tibco [Software Inc./TIBX] Spotfire is a valid offer but not used by our clients."
- "The best MDM platform is Informatica for clients in heterogeneous environments, but SAP and Oracle middleware and connectors will be the main solutions adopted in their own client bases."
- "All companies with innovative products are benefitting. I am looking especially at vendors that use Hadoop software and most of all predictive analytics."

The best MDM platform is Informatica for clients in heterogeneous environments, but SAP and Oracle middleware and connectors will be the main solutions adopted in their own client bases.

BI Specialist
EMEA Integrator

➤ **Head of the middleware practice with an EMEA IT consulting and outsourcing company**

With an abundance of competition and specific needs, the market for big data management is becoming increasingly fragmented, prompting companies to spend more time evaluating their options. Informatica is a leader in big data and one of several companies used most commonly by this consultancy's clients. Informatica serves all industries, is more agile than Oracle and IBM and helps companies reduce cost while increasing operational capacity. Oracle struggled with bringing its Fusion solution to market too slowly. SAP and Oracle provide an end-to-end solution for their clients, but emerging companies are creating greater competition for all providers.

- "The market is becoming more fragmented in the new areas due to all the competition and the rise of new independent companies, which makes it harder for any one company to have an advantage."
- "There are many different platforms according to the vertical the company is in and their strategy. Tibco is used a lot in financial services, and Informatica is used across all industries. Right now, companies are still evaluating which software to use and slowly are going ahead with their plans."
- "Oracle is penalized because they are slow with Fusion while Software AG [SOW] and IBM are doing well. Competition is definitely increasing. On the service-oriented architecture [SOA] we mostly see Tibco, Software AG, Oracle and IBM. For purely managing big data we see Informatica, SAS and IBM."
- "Informatica, SAS and IBM are the most used in our client base. Informatica is well positioned in hot subjects such as big data; they are an agile company compared to the very slow IBM and Oracle. Oracle is still late with Fusion; everybody still is waiting for tons of releases. The integrated vendors' [SAP, Oracle and IBM] strength is that they impose their software. They do agreements with headquarters, and worldwide all branches have to

The market is becoming more fragmented in the new areas due to all the competition and the rise of new independent companies, which makes it harder for any one company to have an advantage.

Head of the Middleware Practice
EMEA IT Consulting Company

implement that specific software. In reality, if you let everybody choose, you will see many different choices with all the independent solutions you can find today."

- "Tibco and Software AG are getting implemented quite a bit in our client base for data analysis. Informatica is playing the card of modernization and information life cycle management [ILM], which can reduce costs and increase operational capacity."
- "Oracle IBM and SAP react by proposing integrated solutions A to Z to all their clients. IBM has the integrated WebSphere platform and Oracle the Fusion middleware, which is not completely ready. All solutions will have to be ready and operational if they want to succeed. The crucial points will be SOA and data management since the other tools are ready and working. However, there are many small emerging vendors now which is why this market is very interesting."
- "The most used MDM platforms are Informatica, Tibco and SAP according to what the client uses for ERP."

> Informatica, SAS and IBM are the most used in our client base. Informatica is well positioned in hot subjects such as big data; they are an agile company compared to the very slow IBM and Oracle.
>
> *Head of the Middleware Practice*
> *EMEA IT Consulting Company*

- **David Douglas, co-founder of CrinLogic, a big data consultancy**

The big data solution of choice is the Apache Hadoop ecosystem of open source products, including packaged Hadoop solutions from Hortonworks Inc., Cloudera Inc. and 10gen Inc.'s MongoDB. Companies are just beginning to experiment with big data solutions, and the attraction of Hadoop is lower cost and proven scalability by early adopters. The source recommends open source because Oracle, IBM, SAP and others are less advanced. Big data is a young market with few skilled professionals and as yet lacks middleware winners and losers. The source knew of no company currently using SAP HANA.

- "There is widespread confusion about what big data is. Many still confuse data size as the only entry criterion and neglect type of data (unstructured and structured) and data velocity. We see this as being a normal problem consistent with the early adoption phase of big data. This is confounded, however, by all the competing vendor products, many of which oversell the true capabilities of their systems. We are just beginning to understand how to leverage big data and the solutions market for products is just beginning to get really interesting."
- "Our clientele, which is outside the 'early' adopter community of social media, online retail, and certain government agencies, is in the experimentation phase of big data 'solutioning.' They are experimenting largely with the Apache Hadoop ecosystem of products for data store [Cassandra, HBase, HDFS], development tool [MapReduce, Pig], and analysis [Hive, Mahout]."
- "The attraction of these tools is multifold: 1) Open source has an attractive cost structure; 2) there are lower hardware costs and it runs on commodity hardware; and 3) they are proven scalable in companies such as Facebook, Yahoo, and LinkedIn."
- "We continue to recommend the open source route for big data solutions. It is our belief that the large players in the traditional RDBMS [relational database management systems] market such as Oracle, IBM, and EMC are still playing catch-up. Their most recent plays have been to partner with leaders in the open-source big-data market."
- "From a data management perspective, our choice remains the Apache Hadoop ecosystem of projects. We believe there is value in the packaged solutions of Hadoop offered by both Hortonworks and Cloudera. We also recommend MongoDB depending on the specific requirements of our customers. The front-end analytics side is quite dynamic at this stage. For companies with sophisticated analytic capabilities, we generally recommend combining various tools such as R, SAS or Mahout."

> We continue to recommend the open source route for big data solutions. It is our belief that the large players in the traditional RDBMS market such as Oracle, IBM, and EMC are still playing catch-up. Their most recent plays have been to partner with leaders in the open-source big-data market.
>
> *Co-founder of CrinLogic,*
> *Big Data Consultancy*

- "For companies possessing the more traditional analytic capabilities one may find in a business intelligence setting, there are a host of useful tools such as Informatica, Hive, Karmasphere [Inc.] and Datameer [Inc.]."
- "For big data, the most widely adopted data management platform is Apache Hadoop. Regarding platforms for data analysis this is still an immature market. I do not believe there is any leader per se. Big data analytics currently require sophisticated data scientist skills that are rare. These types of individuals tend to use SAS and

R with big data. I do not foresee that changing anytime soon. Rather, I foresee SAS and R integrating further with Hadoop. Machine learning tools and data visualization tools will become highly utilized in the big data space. There are no current winners here yet though."

- "Apache Hadoop projects will continue to dominate the market with the help of companies such as Cloudera and Hortonworks. MongoDB, CloudDB, MapR and Hadept will also see a lot of traction. One area where there is a lot of interest is near real-time and real-time analysis tools and techniques."
- "Although I wouldn't say they are falling behind per se, it is clear that the traditional data management companies such as Oracle, IBM and Microsoft are still trying to figure out this space."
- "I do not know of any customers using HANA at this time. I cannot rate the middleware players because this is still an evolving market. It is still up in the air, too early to tell who has the best or worst products."

> Although I wouldn't say they are falling behind per se, it is clear that the traditional data management companies such as Oracle, IBM and Microsoft are still trying to figure out this space.
>
> *Co-founder of CrinLogic, Big Data Consultancy*

3) INDUSTRY SPECIALISTS

These four sources reported seeing considerable interest in big data but a limited number of concrete projects because companies still must internalize the use of big data in their business processes. The market is fragmented based on the various needs and uses of analytical tools for managing big data, giving rise to many potential players and no overwhelming favorite. Social listening and Web data mining are done mainly in the Hadoop environment where they are then made available for companies to use. Some of this data is cleaned and sorted either with search engines such as HP Autonomy and Dassault Exalead or with data quality management (DQM) software such as Informatica, which was noted as a leader in this space because of its complete solution. Data storage leaders include Oracle, IBM and Teradata. In an SAP environment, HANA can do the job with Business Objects ETL tools. Teradata offers a good tool to work with large amount of data while MicroStrategy can act as the BI tool to handle large volume.

➤ Business intelligence analyst for a technology consulting and benchmarking company

Big data projects are slow to develop as the technology is new and evaluations continue to take place on the best solutions. Hadoop software is up-and-coming with the ability to clean and organize massive amounts of data. SAP, Oracle, Informatica and IBM also offer strong solutions, but a leader has yet to emerge. SAP tools will be most commonly used as connectors for HANA.

- "Right now we don't know what the best option is. Managing big data is a new area. Companies have two different sets of data, the structured ones and the ones that come from social media, the cloud, the Web, unstructured or semistructured. All that data needs to be reconciled, sorted, kept or eliminated."
- "Companies still don't know how to work with big data in real time. They need to change their internal processes before they can fully take advantage of this. It is changing, but it is a minority. In the conference I attended today, when they asked who had big data projects going, only three people raised their hands."
- "First, you need to clean it and sort it. For example, 85% of the data from Twitter has to go. Only the pertinent data is kept. You can't pollute the enterprise with all that data that most of the time is irrelevant. Right now, to do that, there is some different software based on Hadoop."
- "You need to have different tool families—audio miner, text miner, log analyzer, Web crawler, Web harvester, profiling, ad server and so on—so that you can analyze audio and video files, which can be important. To extract the data from the Web, the first ETL entirely written for Hadoop is Hurence, a relative unknown. However, most ETL vendors do the job in some way, and the top ones are the usual: Informatica, IBM, Oracle and SAP."

> Companies still don't know how to work with big data in real time. They need to change their internal processes before they can fully take advantage of this. It is changing, but it is a minority. In the conference I attended today, when they asked who had big data projects going, only three people raised their hands.
>
> *Business Intelligence Analyst Technology Consulting Company*

- "There are different data platforms. You can use in-memory databases such as HANA and then BI tools. But the data you have needs to be cleaned and treated with a data quality management software such as Informatica and others."
- "SAP, Informatica, Oracle and IBM are the known leaders for data analysis, but for big data specific it is too early to tell who'll be the best."
- "Technology based on Hadoop and Web search engines are the up-and-comers. With a combination of data mining software and search engine, you can produce some clean data. Statistically you only need a sample of 10,000 to be good. With 10,000 entries, you can only be 1% off in your analysis. The new technology also has to produce data to work with BI software. Unfortunately, all those new tools I mentioned before for data mining are not easy to use, and companies will not find resources easily as there aren't any. At the moment those new tools are not being used too much."
- "Forty-five percent of the users of big data are marketing and advertising people using BI software. The rest are different industries such as telco companies or companies that have to capture large amounts of data mainly from consumers. You would think that projects such as IBM Smarter Cities would be the ones generating and using most of big data, but in fact it is still consumer-oriented companies who do the most."
- "The leaders are still the same as before and the big data tools are still in process, but I don't know who the best is yet. It's also too early to say who will fall behind and be losers in big data."
- "SAP will provide most of HANA connectors. The BOBJ data services are working well enough for SAP clients."
- "All the vendors who are good on Hadoop and provide the best ETL and even an integrated tool for big data will be well positioned to win."

> **Technology based on Hadoop and Web search engines are the up-and-comers. With a combination of data mining software and search engine, you can produce some clean data.**
>
> *Business Intelligence Analyst*
> *Technology Consulting Company*

- Business intelligence expert for an IT consulting company

The market for big data is growing with projects on the horizon, but adoption currently is slow. Many competitors offer tools for specific uses, resulting in numerous companies gaining business. Informatica is a leader with its MDM platform, which is a complete solution ahead of others. EMC and HP excel at storing the data. Meanwhile, SAP's connectors will get the lion's share of the work with SAP clients and HANA, which is gaining a following and positioning SAP to grow.

- "All that market is in front of us. There is not massive adoption right now. Right now, companies have classical solutions: BI software to analyze structured data coming from ETL tools. For nonstructured data they use intelligent search engines."
- "Informatica's MDM platform is performing well. They are ahead in general and have a complete solution with an excellent team. For SAP users with 100% SAP environment, SAP MDM is fine."
- "For storing and working on big data, the solutions I have seen the most are EMC Greenplum, HP Vertica with Autonomy, Exalead and Sinequa."
- "For ETL, I have not seen the leader yet. Informatica is strong and you have other ETL from SAP and IBM, but nobody is marketing a big data ETL really. They might do the job, but they were not created for this. We have seen Ab Initio also being adopted in large companies with success."
- "HANA with BOBJ is generating a lot of interest as a leading platform for data analysis, and it looks promising."
- "SAP will provide most of HANA's connectors, already the BOBJ data services are working well enough, for 100% SAP clients."
- "SAP will gain thanks to HANA's power."

> **Informatica's MDM platform is performing well. They are ahead in general and have a complete solution with an excellent team. For SAP users with 100% SAP environment, SAP MDM is fine.**
>
> *Business Intelligence Expert*
> *IT Consulting Company*

- Middleware expert for a worldwide IT consulting company

Many different options exist in managing big data. Leaders emerge depending on the type of data and what a company intends to do with it. Hadoop is the leader in mining social data. Informatica and IBM lead in working with structured data. SAP improved its offerings with the 2008 acquisition of Business Objects. Working on the data in the cloud is a valued option, particularly for this source's analysis of social data.

- "Everything is moving right now, but there are different options according to what kind of data. If you are doing social listening and have unstructured data, the best option is a Hadoop platform with the related data mining software. You can clean the data there and send it to a search engine or a BI software to analyze, even with Excel. With structured data, the best tools are Informatica and IBM and then it depends on what ERP you have, what BI you have. SAP has improved a lot after the BOBJ acquisition with the data service ETL that is included in BOBJ."
- "Similarly for structured data you can use different platforms such as HANA, Teradata or Ab Initio according to what vertical you are in and what problems you have. Teradata is good with big volume; others are good in some configuration. An SAP client will want HANA."
- "All the technology around Hadoop is good. The best option is to have all this in the cloud and in the cloud work on the data, clean the data and have it ready to be analyzed there. We are currently evaluating to move our social listening platform to the cloud."
- "All SAP clients' best option is usually SAP ETL, but if they need data from the Web, I am not sure they can connect it to HANA. In that case, they can outsource the job to companies who have an Hadoop platform and can deliver clean data to them to work with whatever software they have."

> **If you are doing social listening and have unstructured data, the best option is a Hadoop platform with the related data mining software. You can clean the data there and send it to a search engine or a BI software to analyze, even with Excel. With structured data, the best tools are Informatica and IBM.**
>
> *Middleware Expert*
> *Worldwide IT Consulting Company*

- **Veteran business intelligence thought leader, consultant, author and speaker**

  SAP has its own business intelligence tools optimized for and tailored to HANA. The source reserved judgment on HANA's potential but pointed out that SAP has 300 new customers. HANA may eliminate Oracle's grasp on SAP customers, but the source doubted HANA will eliminate the need for storage because companies still will require a disk-based data warehouse to store historical perspectives. Oracle's Exadata and IBM's Netezza are selling well. Hadoop has the biggest mindshare, but its future is unclear.

  - "SAP is optimizing Business Objects tools for HANA, providing access to data structures in HANA that other tools won't be able to access and via special APIs. The same was true with their predecessor tools, BEx. The SAP tools should work best with HANA. Oracle and IBM have their own BI tools, but Oracle is not necessarily optimizing them for Exadata. Teradata doesn't have BI tools."
  - "[SAP's] Sybase IQ has been out there for 15 years, and it has a ton of customers. The question is whether Sybase will retain that base of customers and grow their market share. The market is more competitive now. They also have deeper pockets with SAP, and I've noticed they are a lot more visible now. SAP is banking a lot of its company on HANA. They have pointed to 300 new customers for HANA and they have been talking it up, but it is still a young technology."
  - "Oracle Exadata is slightly different than the other appliances because it handles transactions and can be tuned by customers. It is doing very well and selling strongly. IBM is doing well with Netezza, and Netezza was doing well even before IBM bought them."
  - "If HANA can do everything in-memory that would be fabulous, but most will tell you big data and in-memory don't necessarily go well together. You can't put tens of terabytes in-memory. You can have a terabyte in-memory, but eventually you need to store it. BW is just another app that runs on HANA, one of the first that SAP will release actually."
  - "SAP's plan is to kick Oracle out of their accounts, and [HANA] will help to do a good bit of that. I think this was the missing piece in [SAP's] portfolio, not to underestimate how important performance is. We will see."
  - "Hadoop has a lot of mindshare right now. It is open source, cheaper, and developers generally don't want to deal with the expense or SQL development using relational databases when processing unstructured log data. A

> **SAP's plan is to kick Oracle out of their accounts, and [HANA] will help to do a good bit of that. I think this was the missing piece in [SAP's] portfolio, not to underestimate how important performance is. We will see.**
>
> *Veteran Business Consultant*

lot of MySQL developers are jumping on the Hadoop bandwagon when they encounter big data problems that MySQL can't handle."

- "Hadoop is just the newest kid on the block. But for those who want to do a lot of comparing and exploring, they are playing with it to see where it fits and what it can do. It seems it can do everything they want for a lower licensing cost. But some are still trying to figure out what it is best suited for."
- "SAP does have a whole bunch of data integration suites that came over as part of the SAP acquisition of Business Objects. SAP also has Sybase Replication Server, which also populates HANA."
- "Informatica's main product is PowerCenter. IBM Cognos and Qlik Technologies are front-end BI tools."

4) DATABASE CONSULTANTS

The field is crowded with quality offerings from IBM, SAP, Oracle, HP, EMC and Teradata. Teradata together with MicroStrategy and Informatica are best-of-breed and liked because they are independent. SAP and Oracle tend to sell mainly to their own clients, but SAP is gaining market share because its clients with an Oracle database are considering a database change in light of real-time computing and in-memory solutions. SAP's HANA is gaining traction from successful proofs of concept. The midmarket has no clear leader as the big appliances are too expensive and companies tend to go with solutions in the cloud and open source software such as Hadoop.

➤ Database expert with a large consulting company

The market lacks a clear leader offering a solution for all needs. Companies using SAP or Oracle will stay with those providers for big data management tools. Informatica, Tibco and Hadoop offer quality solutions and operate independently, endearing themselves to non-Oracle or non-SAP clients. This source's clients are beginning to consider changing databases, a marked difference from a year ago. Oracle is the most threatened because its database offering with its large installed base is deemed insufficient, giving rise to possible defections to HANA or Hadoop-based solutions.

- "If you take the midmarket, there is no leader and companies use all sorts of different solutions. This is an untapped market with huge growth opportunities. All software around Hadoop, open source and Microsoft is being evaluated, but there is no clear winner yet."

> “Clients are ready to talk about changing database or at least work with more than one database. That was not true last year. … The company most at risk is Oracle as they have sold plenty of databases to those companies and now those databases are not enough to handle the kind of data we have today.
>
> *Database Expert*
> *Large Consulting Company*

- "Informatica has good tools of data quality. Hadoop connectors and Tibco are able to work very fast. They are leaders in the data processing and are still independent, so many companies like them. They are superior to Oracle and SAP, who limit themselves to their own data. HANA does not work with Oracle database well, and Oracle is only fast if you work with Oracle data in their own appliance."
- "Usually there is an Hadoop environment with nonstructured data, one relational database with structured data and some mobile database such as Sybase. The Hadoop environment sometimes is separated and sometimes feeds unstructured data revisited to the structured database to cross-reference, for example, comments from Twitter to match an entry in CRM data in-house. That requires DQM tools like Informatica."
- "Clients are ready to talk about changing database or at least work with more than one database. That was not true last year. Already with SAP you work with three databases now: Sybase, Oracle or DB2, and HANA."
- "The top offers are IBM Stream computing software and Software AG's Terracotta. Both load data in real time. There are also plenty of other good offers with Teradata and MicroStrategy, HP, EMC, and those are mostly appliances to handle big data for large companies. It's difficult to understand which one is the best."
- "Big appliances will not be successful here so we have to watch carefully who will find the best solution to handle big data in a cloud or server grids. The company most at risk is Oracle as they have sold plenty of databases to those companies and now those databases are not enough to handle the kind of data we have today."
- "Working with big data is an opportunity and a reason to think about database strategy and Oracle. If companies decide to keep Oracle, they will also have to have some in-memory databases. Oracle will be one of the

databases used, not the only one. Some companies might also decide to remove Oracle and use different in-memory databases for different applications. One case can be HANA plus Sybase instead of Oracle, or some solutions based on Cloudera and Hadoop as well as Microsoft and MySQL. We expect big changes in the database market in the next two years."

- "There are many multinational projects in the pipeline, mostly coming from the BI area. A good number of clients started to think seriously about real-time computing and especially SAP HANA. HANA is not a mature offer yet, but they are getting some traction."
- "The area where we see more talks is at the database and data level. Many times, with big data the usual relational databases do not perform well enough. And with the requirements of real-time computing, data batch processing is not enough. With HANA or real-time computing, you can't have slow access to data. And big data needs to be processed fast. We can now see that many different environments are required and coexistent."

➤ Oracle partner with a large consulting company

As an Oracle partner, the majority of this source's clients are using Oracle tools for data management though big data projects have slowed. SAP clients are likely to stay with SAP tools. HANA is getting good word of mouth while in the proof-of-concept stage.

- "Our clients talk a lot about big data. However, market demand is flat right now as we don't see many big projects but mostly harmonization and rationalization of the existing systems. Most high-end projects are fueled by an upgrade or work on data warehouse or analytics. Several projects are around HCM [human capital management] where there is less saturation."
- "Our Oracle clients tend to buy mostly Oracle tools. Oracle middleware is very good and all the tools around Hyperion tend to be technologically good—for example, all EPM [enterprise performance management] and ETL. So in general we can say that Oracle clients stay with Oracle and SAP clients stay with SAP. SAP is starting some good projects around HANA, and I heard there is a lot of interest even if they are still in proofs of concept. In that case, I do not know if clients would maintain Oracle database or change to HANA."
- "We haven't really seen any Exalytics implementations so far. We have had no requests from our clients yet, so it is difficult for me to have any opinion about a part that is an expensive box."

> SAP is starting some good projects around HANA, and I heard there is a lot of interest even if they are still in proofs of concept. In that case, I do not know if clients would maintain Oracle database or change to HANA.
>
> *Oracle Partner*
> *Large Consulting Company*

➤ CEO at a Sybase reseller and consultancy in the EMEA

SAP's Sybase favors structured data and works well with HANA, Oracle, IBM and Hadoop to organize data on mobile devices.

- "Sybase does not really work with unstructured data; it was made to work with structured data. Sybase SQL Anywhere is used for several reasons, such as database server for work groups or for small or medium-sized businesses."
- "Its best use is as a mobile database as it includes scalable data synchronization technology that provides change-based replication between separate databases including Oracle and IBM DB2. With Mobilink SQL Anywhere can get some unstructured data and a connector exists for the Hadoop framework."
- "However, Sybase's strength is the ability to bring all the data, structured or unstructured but organized, on mobile devices. Sybase can work with HANA to bring a ton of data to users' fingertips on their mobile devices."

Secondary Sources

Eight secondary sources discussed the broader big data market, growth of Hadoop, analytical processing systems, two wins for Informatica, an SAP application allowing access to HANA from the iPad, and Teradata's challenge to Oracle's Exalytics.

- Oct. 18, 2011, Information Week article
  Twelve top big-data players are profiled in a slideshow with details on each company, their products and the role they play in the big data environment.
  - "This image gallery presents a 2011 update on what's available, with options including EMC's Greenplum appliance, Hadoop and MapReduce, HP's recently acquired Vertica platform, IBM's separate DB2-based Smart Analytic System and Netezza offerings, and Microsoft's Parallel Data Warehouse. Smaller, niche database players include Infobright, Kognitio and ParAccel. Teradata reigns at the top of the market, picking off high-end defectors from industry giant Oracle. SAP's Sybase unit continues to evolve Sybase IQ, the original column-store database."

- March 27 Business Insider article
  Big data was likened to Twitter in 2008, when the social media company was misunderstood and underestimated, and stands to be very profitable in two years.
  - "In 2008, when Howard Lindzon started StockTwits, no one knew what Twitter was. Obviously, that has changed."
  - "Now that Twitter is more of a mainstream communication channel, Lindzon has figured out the secret to getting past all the noise on Twitter. By using human curation, StockTwits can serve up relevant social media content to major players like MSN Money."
  - "Lindzon said there are three key aspects that have helped solve the spammy nature of Twitter: StockTwits uses humans to curate social media content. The technology filters out penny stock mentions. It has house rules that people must follow or else they get kicked out of it."
  - "It's working: there were 63 million impressions of messages viewed yesterday. This is double from a few months ago."
  - "The value in big data, like the sentiment in tweets, is not yet understood, Lindzon said. Just like the value of Twitter as a communication platform was misunderstood in 2008."
  - "'Prices and business models are being made up now because this data is so fresh and interesting and real time. In 2014 people will say wow—that's not just interesting, that's wicked profitable.'"

- Feb. 6 BeyeNetwork.com blog
  Hadoop and analytical platforms comprise the two markets for big data. This article takes a look at each and compares their different value propositions.
  - "There are two types of Big Data in the market today. There is open source software, centered largely around Hadoop, which eliminates upfront licensing costs for managing and processing large volumes of data. And then there are new analytical engines, including appliances and column stores, which provide significantly higher price-performance than general purpose relational databases. … Both sets of Big Data software deliver higher returns on investment than previous generations of data management technology, but in vastly different ways."
  - "Hadoop is an open source distributed file system available through the Apache Software Foundation that is capable of storing and processing large volumes of data in parallel across a grid of commodity servers. Hadoop emanated from large internet providers, such as Google and Yahoo, who needed a cost-effective way to build search indexes."
  - "Today, many companies are implementing Hadoop software from Apache as well as third party providers, such as Cloudera, Hortonworks, EMC, and IBM. Developers see Hadoop as a cost-effective way to get their arms around large volumes of data that they've never been able to do much with before. For the most part, companies use Hadoop to store, process, and analyze large volumes of Web log data so they can get a better feel for the browsing and shopping behavior of their customers."
  - "Besides being free, the other major advantage of Hadoop software is that it's data agnostic. … Unlike a data warehouse or traditional relational database, Hadoop doesn't require administrators to model or transform data before they load it. … This significantly reduces the cost of preparing data for analysis compared to what happens in a data warehouse. Most experts assert that 60% to 80% of the cost of building a data warehouse, which can run into the tens of millions of dollars, involves extracting, transforming, and loading (ETL) data. Hadoop virtually eliminates this cost."

> **Many companies are starting to use Hadoop as a general purpose staging area and archive for all their data.**
>
> *BeyeNetwork.com Blog*

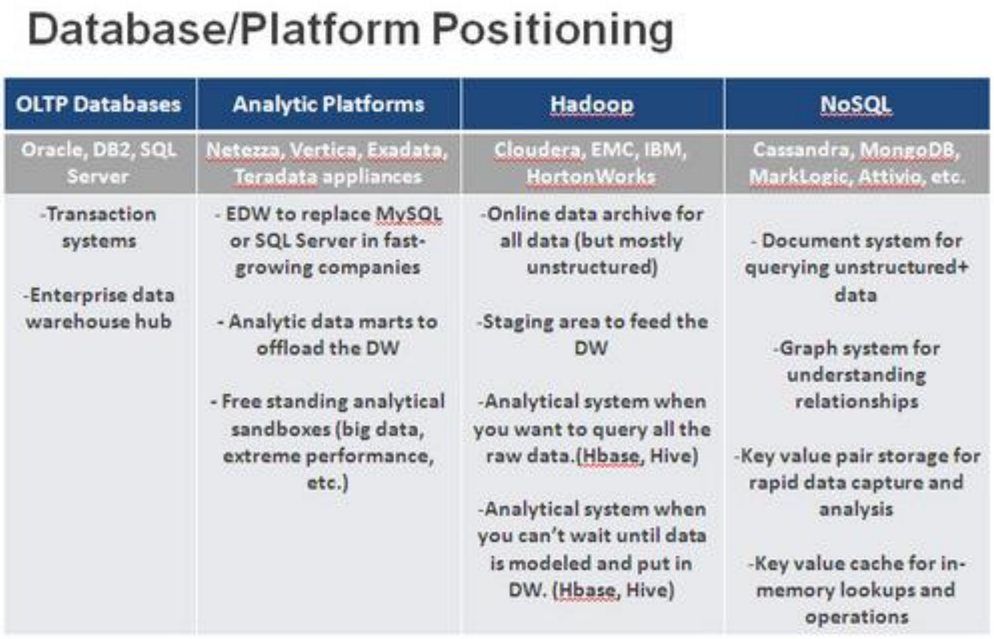
- "As a result, many companies are starting to use Hadoop as a general purpose staging area and archive for all their data. So, a telecommunications company can store 12 months of call detail records instead of aggregating that data in the data warehouse and rolling the details to offline storage. With Hadoop, they can keep all their data online and eliminate the cost of data archival systems. They can also let power users query Hadoop data directly if they want to access the raw data or can't wait for the aggregates to be loaded into the data warehouse."
- "Of course, nothing in technology is ever free. When it comes to processing data, you either 'pay the piper' upfront, as in the data warehousing world, or at query time, as in the Hadoop world. … So a Hadoop developer ends up playing the role of a data warehousing developer at query time, interrogating the data and making sure it's format and content match their expectations."
- "But what's more costly is the expertise and software required to administer Hadoop and manage grids of commodity servers. Hadoop is still bleeding edge technology and few people have the skills or experience to run it efficiently in a production environment. … Hadoop's latest release is equivalent to version 1.0 software, so even the experts have a lot to learn since the technology is evolving at a rapid pace."
- "The other type of Big Data predates Hadoop and NoSQL variants by several years. This version of Big Data is less a 'movement' than an extension of existing relational database technology optimized for query processing. These analytical platforms span a range of technology, from appliances and columnar databases to shared nothing, massively parallel processing databases. The common thread among them is that most are read-only environments that deliver exceptional price-performance compared to general purpose relational databases originally designed to run transaction processing applications."
- "Although the pricetag of these systems often exceeds a million dollars, customers find that the exceptional price-performance delivers significant business value, in both tangible and intangible form. For example, XO Communications recovered $3 million in lost revenue from a new revenue assurance application it built on an analytical appliance, even before it had paid for the system! It subsequently built or migrated a dozen applications to run on the new purpose-built system, testifying to its value."
- "Kelley Blue Book purchased an analytical appliance to run its data warehouse, which was experiencing performance issues, giving the provider of online automobile valuations a competitive edge. For instance, the new system reduces the time needed to process hundreds of millions of automobile valuations from one week to one day. Kelley Blue Book now uses the system to analyze its Web advertising business and deliver dynamic pricing for its Web ads."
- "First, companies must assess whether an analytical platform outperforms their existing data warehouse database to a degree that warrants migration and retraining costs. … The new analytical platforms usually deliver jaw-dropping performance for most queries tested."
- "Second, companies must choose from more than two dozen analytical platforms on the market today. For instance, they must decide whether to purchase an appliance or a software-only system, a columnar database or an MPP database, or an on-premise system or a Web service. Evaluating these options takes time and many companies create a short-list that doesn't always contain comparable products."
- "Finally, companies must decide what role an analytical platform will play in their data warehousing architectures. Should it serve as the data warehousing platform? If so, does it handle multiple workloads easily or is it a one-trick pony? If the latter, what applications and data sets makes sense to offload to the new system? How do you rationalize having two data warehousing environments instead of one?"
- "Companies that have implemented an enterprise data warehouse on Oracle, Teradata, or IBM often find that the best use of analytical platforms is to sit alongside the data warehouse and offload existing analytical workloads or handle new applications. This architecture helps organizations avoid a costly upgrade to their data warehousing platform, which might easily exceed the cost of purchasing an analytical platform."

> **What's more costly is the expertise and software required to administer Hadoop and manage grids of commodity servers. Hadoop is still bleeding edge technology and few people have the skills or experience to run it efficiently in a production environment. … Hadoop's latest release is equivalent to version 1.0 software, so even the experts have a lot to learn since the technology is evolving at a rapid pace.**
>
> *BeyeNetwork.com Blog*

- Feb. 22 BeyeNetwork.com blog
  Four categories of analytical processing systems are discussed and presented for evaluation, showing the variety of solutions and the possible uses for these solutions.
  - "Faced with an expanding analytical ecosystem, BI managers need to make many technology choices. Perhaps the most difficult involves selecting a data processing system to power a variety of analytical applications."
  - "Instead of selecting a single data management product, BI managers may need to select multiple platforms to outfit an expanding analytical ecosystem. And rather than evaluating four or five alternatives for each platform, the BI manager is faced with dozens of viable options in each category. The once lazy database market is now a beehive of activity!"
  - "Staying abreast of all the new products, partnerships, and technological advances is now a full-time job. Industry analysts who make a living sifting through products in emerging markets are needed now more than ever. Most analysts (including me) will tell you that the first step in selecting an analytical platform is to understand the broad categories of products in the marketplace, and then make finer distinctions from there."
  - "At a high-level, there are four categories of analytical processing systems available today: transactional RDBM systems, analytical platforms, Hadoop distributions, and NoSQL Databases."
  - "Transactional RDBM systems were originally designed to support transaction processing applications although most have been retrofitted with various types of indexes, join paths, and custom SQL bolt-ons to make them more palatable to analytical processing. There are two types of transactional RDBM systems: enterprise and departmental."
    - "**Enterprise Hubs**. The traditional enterprise RDBM systems, such as those from IBM, Oracle, and Sybase, are best suited as data warehousing hubs that feed a variety of downstream, end-user facing systems, but don't handle query traffic directly. Although retrofitted with analytical capabilities, these systems often hit performance and scalability walls when used for query processing along with other workloads and are expensive to upgrade and replace. Thus, many customers now use these "gray-bearded" data warehousing systems as hubs to feed operational data stores, data marts, enterprise reporting systems, analytical sandboxes, and various analytical and transactional applications."
    - "**Departmental Marts**. A number of companies use Microsoft SQL or MySQL as data marts fed by an enterprise data warehouse or as stand-alone data warehouses for a business unit or small- or medium-size business (SMB). Like their enterprise brethren, these systems also often hit the wall when usage, data volumes, or query complexity increases rapidly. A fast-growing business unit or SMB often replaces these transactional RDBM systems with analytic appliances (see below) which provide the same or greater level of simplicity and ease of management as SQL Server or MySQL."
  - "Analytic platforms represent the first wave of Big Data systems. (See "Two Markets for Big Data: Comparing Value Propositions.") These are purpose-built SQL-based system designed to provide superior price-performance for analytical workloads compared to transactional RDBM systems. There are many types of analytic platforms. Most are being used as data warehousing replacements or stand-alone analytical systems."
    - "**MPP Database**. Massively parallel processing (MPP) databases with strong mixed workload utilities make good enterprise data warehouses for analytically minded organizations. Teradata was the first on the block with such a system, but it now has many competitors, including EMC Greenplum and Microsoft's Parallel Data Warehousing Option, which are relative upstarts compared to the 30-year old Teradata."
    - "**Analytical Appliance**. These purposer analyti5iness unit or small

    as those from Teradata, are geared to specific analytical workloads, such as delivering extremely fast performance or managing super large data volumes."

    - "**In-Memory Systems.** If you are looking for raw performance, there is nothing better than a system that lets you put all your data into memory. These systems will soon become more commonplace, thanks to SAP, which is betting its business on HANA, an in-memory database for transactional and analytical processing, and is evangelizing the need for in-memory systems. Another contender in this space is Kognitio. Many RDBM systems are beginning to better exploit memory for caching results and processing queries."
    - "**Columnar.** Columnar databases, such as SAP's Sybase IQ Hewlett Packard's Vertica, Paraccel, Infobright, Exasol, Calpont, and Sand offer fast performance for many types of queries because of the way these systems store and compress data by columns instead of rows. Column storage and processing is fast becoming a RDBM system feature rather than a distinct subcategory of products."
- "Hadoop is an open source software project run within the Apache Foundation for processing data-intensive applications in a distributed environment with built-in parallelism and failover. The most important parts of Hadoop are the Hadoop Distributed File System, which stores data in files on a cluster of servers, and MapReduce, a programming framework for building parallel applications that run on HDFS. The open source community is building numerous additional components to turn Hadoop into an enterprises-caliber, data processing environment. The collection of these components is called a Hadoop distribution. Leading providers of Hadoop distributions include Cloudera, IBM, EMC, Amazon, Hortonworks, and MapR."
- "Today, in most customer installations, Hadoop serves as a staging area and online archive for unstructured and semi-structured data, as well as an analytical sandbox for data scientists who query Hadoop files directly before the data is aggregated or loaded into the data warehouse. But this could change. Hadoop will play an increasingly important role in the analytical ecosystem at most companies, either working in concert with an enterprise DW or assuming most of its duties."



Database/Platform Positioning

OLTP Databases	Analytic Platforms	Hadoop	NoSQL
Oracle, DB2, SQL Server	Netezza, Vertica, Exadata, Teradata appliances	Cloudera, EMC, IBM, HortonWorks	Cassandra, MongoDB, MarkLogic, Attivio, etc.
-Transaction systems -Enterprise data warehouse hub	- EDW to replace MySQL or SQL Server in fast-growing companies - Analytic data marts to offload the DW - Free standing analytical sandboxes (big data, extreme performance, etc.)	-Online data archive for all data (but mostly unstructured) -Staging area to feed the DW -Analytical system when you want to query all the raw data.(Hbase, Hive) -Analytical system when you can't wait until data is modeled and put in DW. (Hbase, Hive)	- Document system for querying unstructured+ data -Graph system for understanding relationships -Key value pair storage for rapid data capture and analysis -Key value cache for in-memory lookups and operations

- March 6 BeyeNetwork.com blog
  Informatica announced a partnership with a leading Hadoop distributor, giving the company another avenue for use.
  - "Informatica this week inscribed another notch in its Big Data belt by inking a partnership agreement with MapR, one of the leading Hadoop distributions in the marketplace. The partnership further opens Hadoop to the sizable

market of Informatica developers and provides a visual development environment for creating and running MapReduce jobs."

- "The partnership is fairly standard by Hadoop terms. Informatica can connect to MapR via PowerExchange and apply PowerCenter functions to the extracted data, such as data quality rules, profiling functions, and transformations. Informatica also provides HParser, a visual development environment for parsing and transforming Hadoop data, such as logs, call detail records, and JSON documents. Informatica has already signed similar agreements with Cloudera and HortonWorks."
- "But Informatica and MapR have gone two steps beyond the norm. Because MapR's unique architecture bundles an alternate file system (Network File System) behind industry standard Hadoop interfaces, Informatica has integrated two additional products with MapR: Ultra Messaging and Fast Clone. Ultra Messaging enables Informatica customers to stream data into MapR, while Fast Clone enables them replicate data in bulk. In addition, MapR will bundle the community edition of Informatica's HParser, the first Hadoop distribution to do so."
- "The upshot is that Informatica developers can now leverage a good portion of Informatica's data integration platform with MapR's distribution of Hadoop. Informatica is expected to announce the integration of additional Informatica products with MapR later this spring."
- "The two companies are currently certifying the integration work, which be finalized by end of Q1, 2012."

> The upshot is that Informatica developers can now leverage a good portion of Informatica's data integration platform with MapR's distribution of Hadoop. Informatica is expected to announce the integration of additional Informatica products with MapR later this spring.
>
> *BeyeNetwork.com Blog*

- March 19 InformationWeek article

  Informatica's software is saving eHarmony time in preparing data in Hadoop for loading into a data warehouse.

  - "This is a story about JSON and Ruby. They were spending too much time together in an unrewarding relationship, so sooner or later it had to end."
  - "JSON (Java Script Object Notation) is what eHarmony uses to capture and move data from its various customer-facing Web sites to its back-end systems. When customers seeking love fill out questionnaires about the dating site's advertised '29 dimensions of compatibility,' for example, JSON encapsulates that data and sends it off wherever it's needed. One destination is Voldemort, the highly scalable, distributed NoSQL data store. Another is Solr, the Apache open-source search platform. "
  - "A third destination is Hadoop. That's where eHarmony's matching algorithms do the work of bringing together compatible customer records. And that's where Ruby comes in. You see, eHarmony can't just load JSON-encapsulated data into its SQL-based IBM Netezza data warehouse. It has to transform the object-encapsulated data into nicely structured information that can be loaded into the appropriate columns and rows in Netezza. For more than two years, eHarmony has been using scripts written in Ruby, the popular object-oriented programming language, to process the JSON data and move it into the data warehouse."
  - "Never mind that writing scripts was time-consuming. In addition, each hourly job also took as long as 40 minutes because it had to run on a conventional server rather than in Hadoop's distributed processing environment. eHarmony had people who knew Ruby, so let's just say it was a 'you'll do for now' relationship."
  - "But then eHarmony started getting serious about its long-term data warehousing prospects. Operations were destined to get bigger, according to Grant Parsamyan, director of business intelligence and data warehousing. Enter Informatica and its PowerCenter data-integration platform, which eHarmony was already using to load as much as seven terabytes per day into Netezza from conventional SQL data sources. Ruby was processing roughly 300 gigabytes per day from Hadoop, but Parsamyan says he expects that volume to get four to five times larger. It was clear the Ruby approach could not scale, he says."
  - "Fortunately, Informatica last fall introduced HParser, a product that moves PowerCenter data-parsing capabilities into the Hadoop distributed processing environment. There, the many processors that

> Fortunately, Informatica last fall introduced HParser, a product that moves PowerCenter data-parsing capabilities into the Hadoop distributed processing environment. There, the many processors that work together can handle transformation jobs quickly, just as they do with massive MapReduce computations.
>
> *InformationWeek Article*

work together can handle transformation jobs quickly, just as they do with massive MapReduce computations."

- "Informatica's HParser community edition handles JSON, XML, Omniture (Web analytics data), and log files. Commercial editions are available for documents (Work, Excel, PDF, etc.) and industry-standard file formats (SWIFT, NACHA, HIPAA, HL7, ACORD, EDI X12, and so on). The package also includes a visual, point-and-click studio that eliminates coding. Once the processing is done, PowerCenter can be used to extract the data from Hadoop and move it into the target destination."
- "In tests completed in November, eHarmony proved the advantages of the HParser approach. 'Using a small Hadoop cluster, jobs that took 40 minutes in Ruby can be completed in about 10 minutes,' Parsamyan says. 'More importantly, as data volumes grow, we can just throw more Hadoop nodes at the problem and scale it up as much as we need to.'"
- "Once the HParser approach is in full production, Parsamyan expects to start loading as much as 1 terabyte per day into the data warehouse in short order, and that will enable more analytic measurement of eHarmony's success. The marketing department uses the data warehouse to measure response to its email and banner advertising campaigns. Product development teams use it to study the success of new site features. And the operations team uses the warehouse to study the health of the business, including membership and revenue trends."
- "With data volumes, velocity, and complexity on the rise, practitioners are turning to highly scalable platforms such as Hadoop. HParser is an early example of the type of new tools they'll need to work with the latest Big Data platforms."

➤ March 13 InformationWeek article

SAP added another processing-intensive application to run on HANA and will make it available for use on the iPad in May.

- "SAP has a long list of applications that will benefit from its Hana in-memory technology, and on Tuesday it added a crucial one, SAP BusinessObjects Planning and Consolidation, to the portfolio of apps certified to run on the database. What's more, executives will be able to do their planning from a new iPad app to be introduced by May."
- "Business Planning and Consolidation (BPC) is used by more than 4,000 SAP customers for setting financial and operational performance goals in areas such as sales and production. The app is a centerpiece of SAP's Enterprise Performance Management suite 10.0, but its performance may suffer when planning involves large data sets. The Hana in-memory database, which holds large-scale, detailed data entirely in random-access memory rather than on hard drive disks, is expected to speed query and analysis activities."
- "'Many of our customers view in-memory-enabled planning as a killer application,' Dave Williams, SAP's head of solution marketing for EPM solutions, told *InformationWeek*. 'Planning is logic-processing intensive and it frequently involves querying large data sets and writing information back into the system.'"
- "Running on Hana, BPC will gain up to 21 times faster access to planning data and faster input of what-if scenario-planning data back into the system, Williams said."
- "BPC is based on the Outlooksoft performance management suite SAP acquired in 2007. In the wake of the acquisition, SAP ported a version of the app onto its NetWeaver middleware to make it compatible with SAP applications and infrastructure."
- "About half of current users are on the .Net version of the app, while the other half are on NetWeaver. Only the NetWeaver version of BPC will be compatible with Hana (version 1.0, service pack 3), as SAP Business Warehouse 7.3, SAP's NetWeaver-based data warehouse, is also required. The upgrade is available at no charge through service pack 6 of NetWeaver BPC version 10.0."
- "SAP has no plans to move the .Net version of BPC onto Hana, but Williams said that app is being kept up to data and will soon support

> SAP has a long list of applications that will benefit from its Hana in-memory technology, and on Tuesday it added a crucial one, SAP BusinessObjects Planning and Consolidation, to the portfolio of apps certified to run on the database. What's more, executives will be able to do their planning from a new iPad app to be introduced by May. … Running on Hana, BPC will gain up to 21 times faster access to planning data and faster input of what-if scenario-planning data back into the system, Williams said.
>
> *InformationWeek Article*

Microsoft SQL Server 2012, which incorporates in-memory analysis capabilities."

- "SAP was expected to demonstrate a prototype BPC app for iPad on Tuesday. The app will enable executives to not only review performance data and drill down on exception conditions, it will also enable them to take action by, say, rejecting and updating forecasts and planning assumptions without having to go to a separate desktop application. The iPad app is expected to be available in time for SAP's annual Sapphire event in May."
- "BPC will continue add in-memory performance enhancements, Williams said, but it will do so through "non-disruptive" service packs that will be released on roughly a quarterly basis. One such update will add automated variance analysis, whereby drill paths and deep data tied to the root causes of exception conditions will be prepopulated behind the scenes. This feature is already available in the .Net version of BPC."

➤ March 13 InformationWeek article
Teradata released an enterprise data warehouse platform that will rival Oracle's Exalytics.

- "Teradata has been working on fast data access for years. With last week's release of the Teradata Active Enterprise Data Warehouse (EDW) Platform 6690, the company says it delivers state-of-the-art query performance and a better approach than that offered by rival Oracle's new Exalytics appliance."
- "The vendor's advances in data-access speed in recent years are tied to Teradata Virtual Storage, software that monitors which data is being queried most often and then automatically moves that data to the fastest storage medium available. Before solid state disk (SSD) drives became affordable, Teradata Virtual Storage moved 'cold' (infrequently accessed) data onto the inner tracks of conventional hard drive disks (HDD) and 'hot' (frequently accessed) data onto the outer tracks, where faster rotation delivered quicker data access."
- "Teradata still uses the inner-track/outer-track technique, and it also supports high-density HDDs suitable for archival storage--super cold (very infrequently accessed) data that you nonetheless want accessible online. At the hot end of the storage spectrum, Teradata added super-fast-access SSDs back in 2010. These drives are as much as 18 times faster in data-access speeds than conventional spinning disks."
- "With the 6690, Teradata says there's a wider range of SSD-to-HDD configurations so customers can better tune the platform to their needs. Firms with few fast queries can dial it down to 6% of total capacity on SSDs, while firms with many such queries can crank it up to 25%. Teradata says its latest Virtual Storage software is also that much smarter, with better algorithms for learning what data to store where, with options now ranging from high-density HDDs, to the inner tracks of standard or high-speed drives, to outside tracks, to SSDs."
- "'The system automatically does the data placement, and it operates at the data-block level, not at the [database] table level, so it provides very granular control,' said Scott Gnau, president of Teradata Labs, in an interview with *InformationWeek*. That granular control makes it possible to place 100% of the data needed for timely queries into SSD storage."
- "Teradata's chief rival, Oracle, late last month introduced Exalytics, an appliance aimed at delivering sub-second response times for data-intensive business intelligence (BI) and performance management applications."
- "Exalytics 'adaptive caching' capabilities sound similar to Teradata Virtual Storage management in that the software monitors workloads generated by Oracle Business Intelligence Enterprise Edition-powered dashboards, queries, and analytic applications and automatically moves the hot data from Exadata (or a third-part source) into the memory of the Exalytics appliance. But there's a crucial difference, according to Gnau."
- "'With Exalytics it's all cache, so it's an incremental copy of data,' Gnau said, describing the box as a 'bolt-on Bandaid' that presents incremental storage, heating, and cooling costs. Teradata, in contrast, stores data once in the most appropriate storage option required, so Oracle is 'solving a performance problem that we don't have,' Gnau said."
- "A final 6690 platform upgrade worth mentioning is a move entirely away from 3.5-inch HDDs to smaller 2.5-inch drives. The footprint of each rack remains the same, but the 6690 can pack up to 360 drives (counting 2.5-

'With Exalytics it's all cache, so it's an incremental copy of data,' Gnau said, describing the the box as a 'bolt-on Bandaid' that presents incremental storage, heating, and cooling costs. Teradata, in contrast, stores data once in the most appropriate storage option required, so Oracle is 'solving a performance problem that we don't have.'

InformationWeek Article

inch SSDs) into each box. That means it offers higher storage density, lower power consumption per terabyte, and reduced cooling requirements for the total data warehousing environment as compared to Teradata's older 6680 platform."

- "'It doesn't sound sexy, but data-center space, power consumption, and cooling requirements are always among the top-five concerns when we survey our customers, so it's a big deal,' Gnau said."

Next Steps

Blueshift will follow four different story lines in separate reports during the next few months. We will evaluate HANA, Oracle's Exalytics, Teradata, QlikView and other database solutions and determine which is best positioned to handle big data and real-time computing. Next we will assess which company has the best analytical platform among MicroStrategy, QlikView, Oracle and SAP. We will also determine which has the best applications for big data. Finally, we will look at the progress of and leaders in cloud and Software as a Service (SaaS) offerings.

Additional research by Carolyn Marshall

The Author(s) of this research report certify that all of the views expressed in the report accurately reflect their personal views about any and all of the subject securities and that no part of the Author(s) compensation was, is or will be, directly or indirectly, related to the specific recommendations or views in this report. The Author does not own securities in any of the aforementioned companies.

OTA Financial Group LP has a membership interest in Blueshift Research LLC. OTA LLC, an SEC registered broker dealer subsidiary of OTA Financial Group LP, has both market making and proprietary trading operations on several exchanges and alternative trading systems. The affiliated companies of the OTA Financial Group LP, including OTA LLC, its principals, employees or clients may have an interest in the securities discussed herein, in securities of other issuers in other industries, may provide bids and offers of the subject companies and may act as principal in connection with such transactions. Craig Gordon, the founder of Blueshift, has an investment in OTA Financial Group LP.