

Solace P&C

Building a Big Data platform with the
Hadoop ecosystem

by Gregg Barrett

Acknowledgement

The presentation draws extensively, and focuses on, the work and viewpoints from industry participants including;

- CITO Research
- Diversity Limited
- Forrester Research Inc
- Gartner
- IBM
- IDC
- ITG
- Intel
- MapR
- Ovum
- Planet Cassandra
- TDWI Research
- 451 Research
- Christopher Bienko @ IBM
- Dirk deRoos @ IBM
- John Choi @ IBM
- Marc Andrews @ IBM
- Paul Zikopoulos @ IBM
- Rick Buglio @ IBM
- Krishnan Parasuraman @ IBM
- Thomas Deutsch @ IBM
- David Corrigan @ IBM
- James Giles @ IBM

References are included in-text as well as in the References section at the end of the report.

Introduction

This presentation provides a brief insight into a Big Data platform for Solace P&C using the Hadoop ecosystem.

To this end the presentation will touch on:

- views of the Big Data ecosystem and its components
- an example of a Hadoop cluster
- considerations when selecting a Hadoop distribution
- some of the Hadoop distributions available
- a recommended Hadoop distribution

A Big Data ecosystem

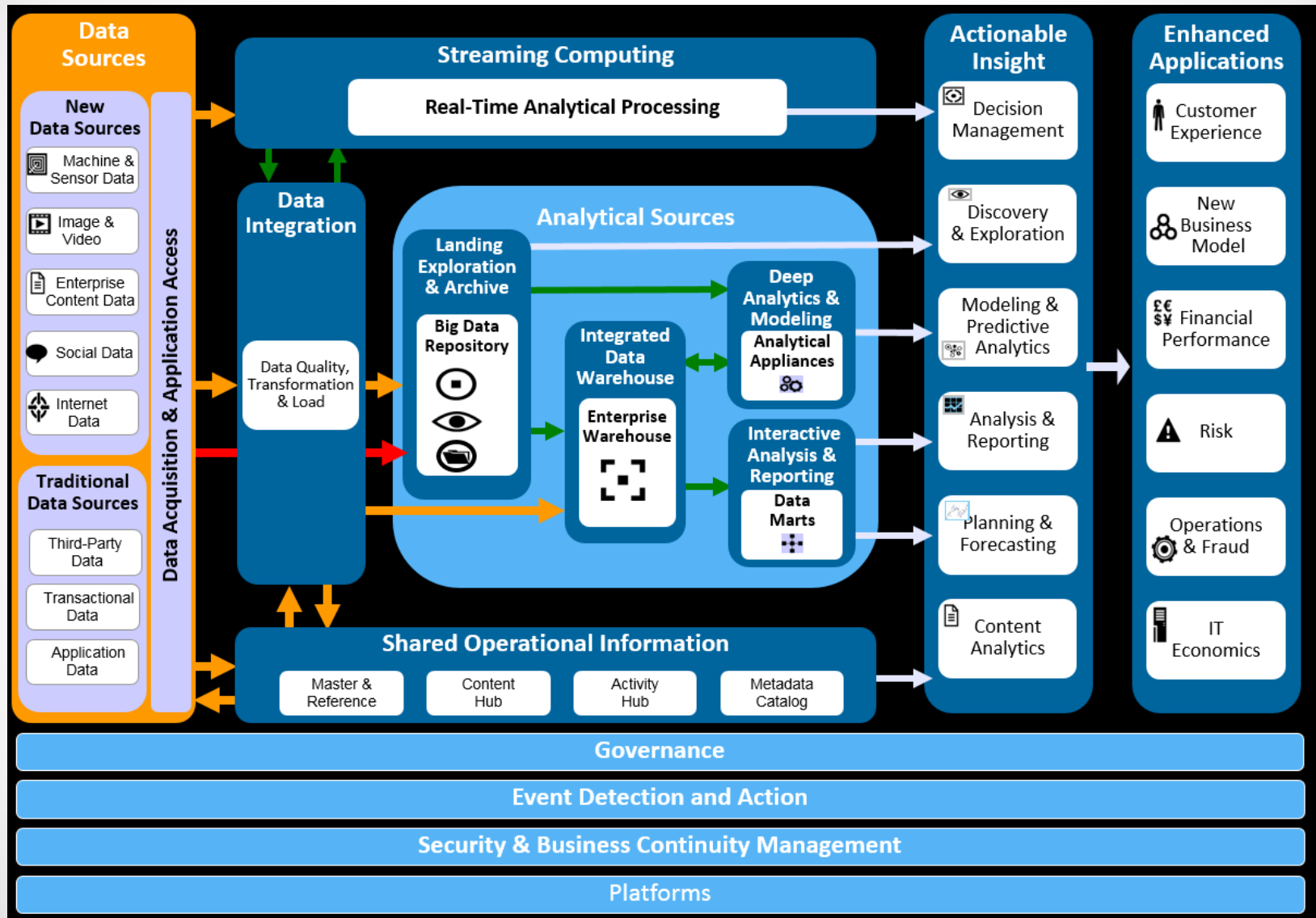


Figure 1. Untitled. Copyright 2015 by IBM. Reprinted with permission.

A Big Data ecosystem with Hadoop

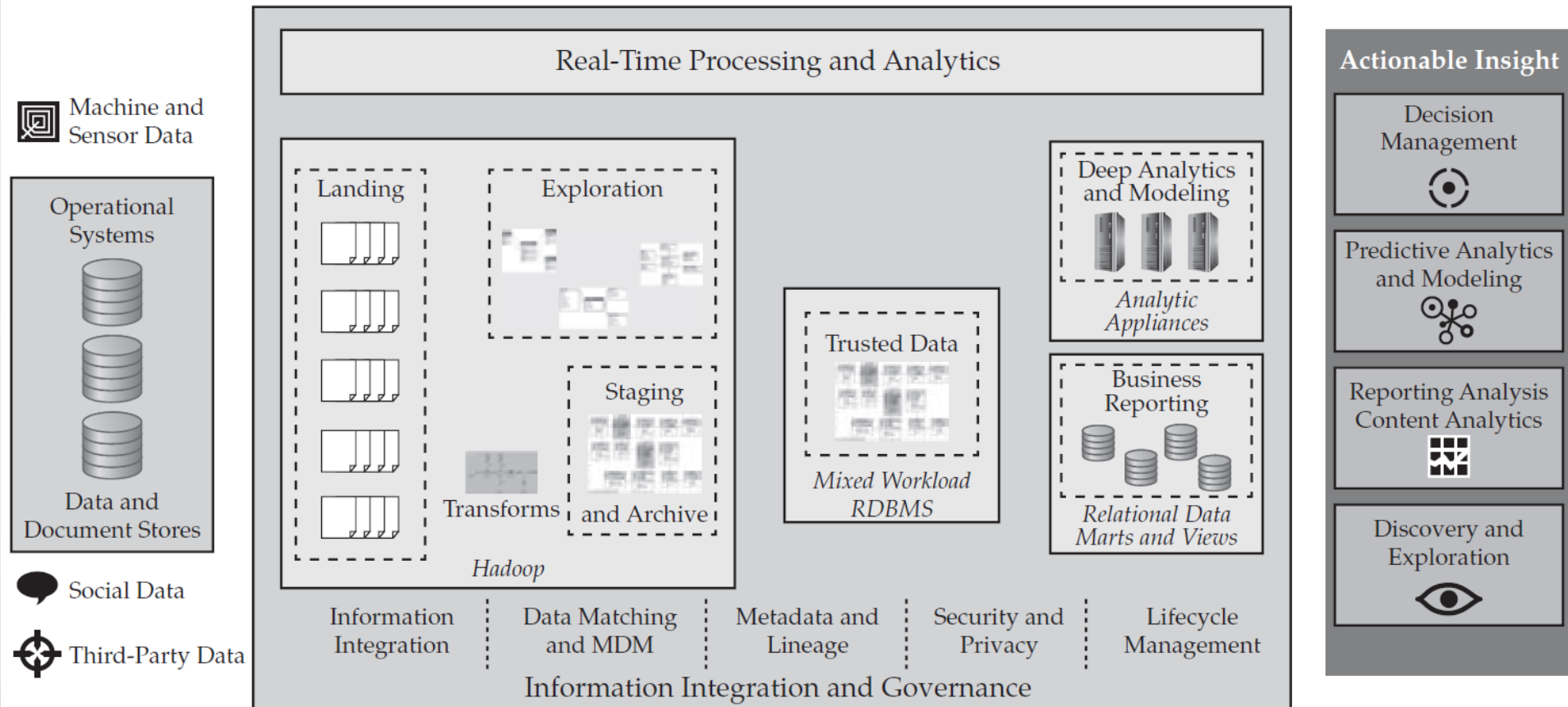


Figure 2. Reference architecture for Big Data and Analytics. Copyright 2014 by McGraw-Hill Education. Reprinted with permission.

Component view of a Big Data ecosystem with Hadoop



Figure 3. Component view of a Big Data ecosystem with Hadoop.

The Hadoop ecosystem

In their book, *Big Data Beyond the Hype*, Zikopoulos, deRoos, Bienko, Buglio and Andrews (2014) classify Hadoop as an ecosystem of software packages that provides a computing framework.

These include:

- MapReduce, which leverages a K/V (key/value) processing framework (don't confuse that with a K/V database);
- a file system (HDFS);
- and many other software packages that support everything from importing and exporting data (Sqoop) to storing transactional data (HBase), orchestration (Avro and ZooKeeper), and more.

When you hear that someone is running a Hadoop cluster, it's likely to mean MapReduce (or some other framework like Spark) running on HDFS, but others will be using HBase (which also runs on HDFS).

On the other hand, NoSQL refers to non-RDBMS SQL database solutions such as HBase, Cassandra, MongoDB, Riak, and CouchDB, among others.

(Zikopoulos, deRoos, Bienko, Buglio, Andrews, 2014, pg. 38)

Key components of many Hadoop environments

MapReduce

MapReduce is a system for parallel processing of large data sets.

According to IBM (2015) as an analogy, you can think of map and reduce tasks as the way a census was conducted in Roman times, where the census bureau would dispatch its people to each city in the empire. Each census taker in each city would be tasked to count the number of people in that city and then return their results to the capital city. At the capital, the results from each city would be reduced to a single count (sum of all cities) to determine the overall population of the empire. This mapping of people to cities, in parallel, and then combining the results (reducing) is much more efficient than sending a single person to count every person in the empire in a serial fashion. (IBM, 2015)

Hadoop

MapReduce is the heart of Hadoop. Hadoop is an open source software stack that runs on a cluster of machines. Hadoop provides distributed storage and distributed processing for very large data sets.

NoSQL

NoSQL is a database environment. Using the definition from Planet Cassandra (2015), a NoSQL database environment is, simply put, a non-relational and largely distributed database system that enables rapid, ad-hoc organization and analysis of extremely high-volume, disparate data types. NoSQL databases were developed in response to the sheer volume of data being generated, stored and analyzed by modern users (user-generated data) and their applications (machine-generated data).

(Planet Cassandra, 2015)

Spark

What is Spark and what does it mean for Hadoop?

IBM (2014) refers to Spark as an open source engine for fast, large-scale data processing that can be used with Hadoop, boasting speeds up to 100 times faster than Hadoop MapReduce in memory, or 10 times faster on disk. As with the early enthusiasm around Hadoop, Spark should not be thought of as a singular platform for analytics, as it can be used with existing investments for the widest variety of data types and analytics workloads. (IBM, 2014)

A closer look: NoSQL and SQL

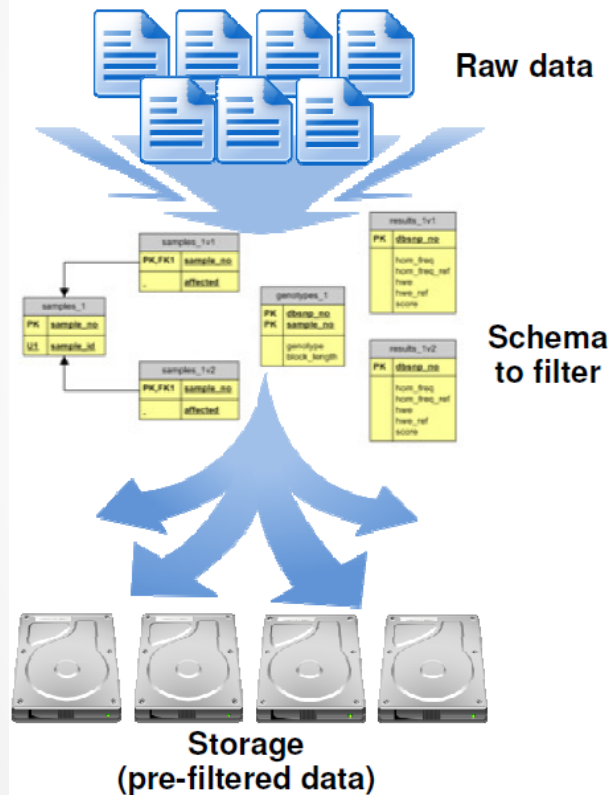
The NoSQL World (Schema Later)	The SQL World (Schema First)
New questions? No schema change	New questions? Schema change
Schema change in minutes, if not seconds	Schema: permission process (we're not talking minutes here)
Tolerate chaos for new insights and agility (the cost of "getting it wrong" is low)	Control freaks (in many cases for good reasons)
Agility in the name of discovery (easy to change)	Single version of the truth
Developers code integrity (yikes!)	Database manages integrity (consistency from any app that access the database)
Eventual consistency (the data that you read might not be current)	Consistency (can guarantee that the data you are reading is 100 percent current)
All kinds of data	Mostly structured data
Consistency, availability, and partition tolerance (CAP)	Atomicity, consistency, isolated, durable (ACID)

Figure 4. Characteristics of the NoSQL and SQL Worlds . Copyright 2014 by McGraw-Hill Education. Reprinted with permission

A closer look: NoSQL and SQL

▪ Regular database

– Schema on load



▪ Big Data (Hadoop)

– Schema on run

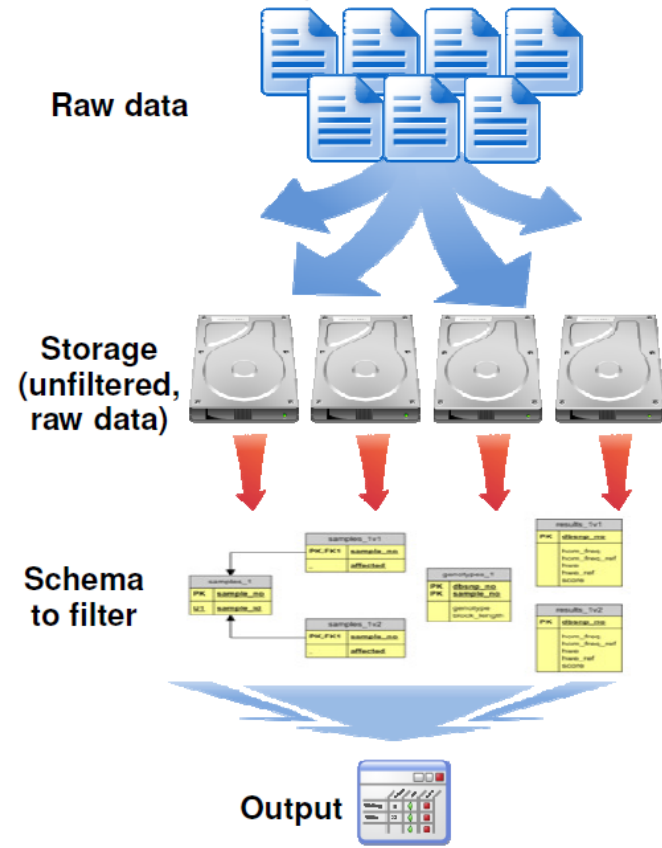


Figure 5. Big Difference: Schema on run. Copyright 2013 by IBM. Reprinted with permission

A closer look: NoSQL and SQL

Conversations around SQL and NoSQL should focus more on how the two technologies complement one another. In the same way that a person living in Canada is better served by speaking English and French, a data architecture that can easily combine NoSQL and SQL techniques offers greater value than the two solutions working orthogonally. In our opinion—and there are those who will debate this—the rise of NoSQL has really been driven by developers and the requirements we outlined earlier in this chapter. New types of apps (online gaming, ad serving, and so on), the social-mobile-cloud phenomenon, and the acceptance of something we call eventual consistency (where data might not necessarily be the most current—no problem if you are counting “Likes” on Facebook but a big deal if you’re looking at your account balance) are all significant factors.

(Zikopoulos, deRoos, Bienko, Buglio, Andrews, 2014, pg. 38)

Hadoop challenges

TDWI Research (2015) in a recent survey found respondents struggling with the following barriers to Hadoop implementation:

Barriers to Hadoop:

- Skills gap
- Weak business support
- Security concerns
- Data management hurdles
- Tool deficiencies
- Containing costs

(TDWI Research, 2015)

According to a study by the International Technology Group, organisations need to be particularly mindful in the highly skilled programming requirements demanded of most Hadoop environments, noting that:

Although the field of players has since expanded to include hundreds of venture capital-funded start-ups, along with established systems and services vendors and large end users, social media businesses continue to control Hadoop. Most of the more than one billion lines of code – more than 90 percent, according to some estimates – in the Apache Hadoop stack has to date been contributed by these.

The priorities of this group have inevitably influenced Hadoop evolution. There tends to be an assumption that Hadoop developers are highly skilled, capable of working with “raw” open source code and configuring software components on a case-by-case basis as needs change. Manual coding is the norm.

Decades of experience have shown that, regardless of which technologies are employed, manual coding offers lower developer productivity and greater potential for errors than more sophisticated techniques. (ITG, 2013, pg. 2)

Hadoop cost implications

Although a Big Data environment such as that illustrated in Figure 3 can be constructed from open source software components, there are still substantial costs involved.

These include:

- Hardware costs
- IT and operational costs in setting up a machine cluster and supporting it
- Cost of personnel to work on the ecosystem

These costs are NOT trivial for the following reasons:

- Dealing with cutting edge technology and finding people who know the technology is challenging
- The technology introduces a different programming paradigm, frequently requiring additional training of existing engineering teams
- These technologies are new and still evolving and are not yet mature in the enterprise ecosystem
- The hardware is server grade and large clusters require resources including network administration, security administration, system administration etc., as well as data centre operational costs including electricity, cooling etc.

A Hadoop cluster example

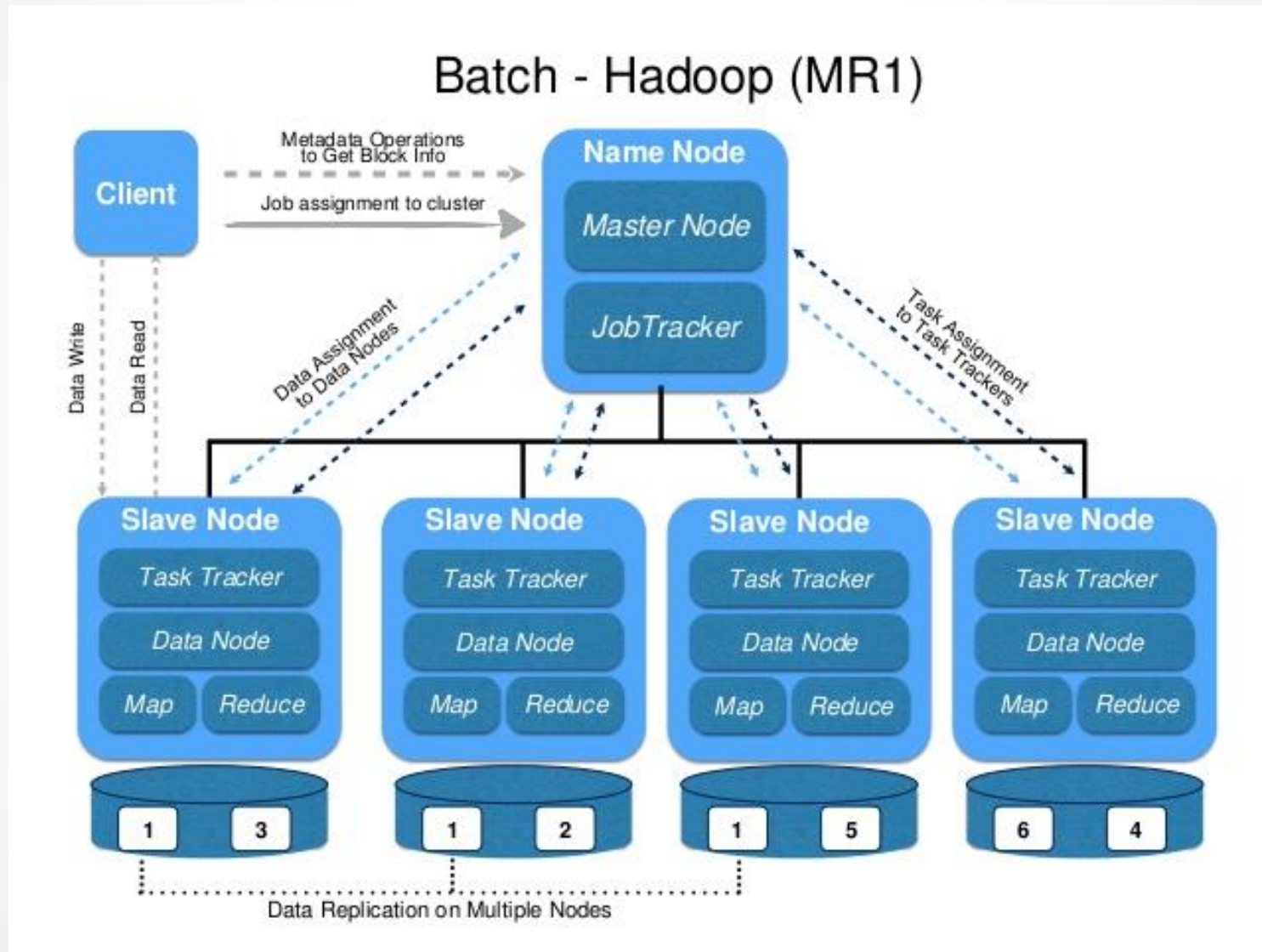


Figure 6. Illustration of a Hadoop cluster. Copyright 2014 by IBM. Reprinted with permission

Hadoop Cluster example

500 TB Hadoop cluster cost calculation example:

- Replication factor = 3
- Size of data to be moved to Hadoop = 500 TB
- Intermediate factor = 1.25 (Hadoop's working space for storing intermediate results of MapReduce jobs.)

Size of cluster: $1\ 875 = 3 \times 500 \times 1.25$

- Using node of 15 TB each

Number of nodes: $125 = 1\ 875 / 15$

- Cost per node = +- 4000 USD

Cost for 125 node Hadoop cluster = 500 000 USD

A Hadoop cluster example

According to MapR (2014) figure 5 below illustrates a cost comparison for a 500 TB cluster between two vendors' Hadoop distributions based on a customer-validated TCO model. The TCO for MapR in the example is \$3.2 million over three years, compared to another Hadoop distribution at \$4.7 million for the same period.

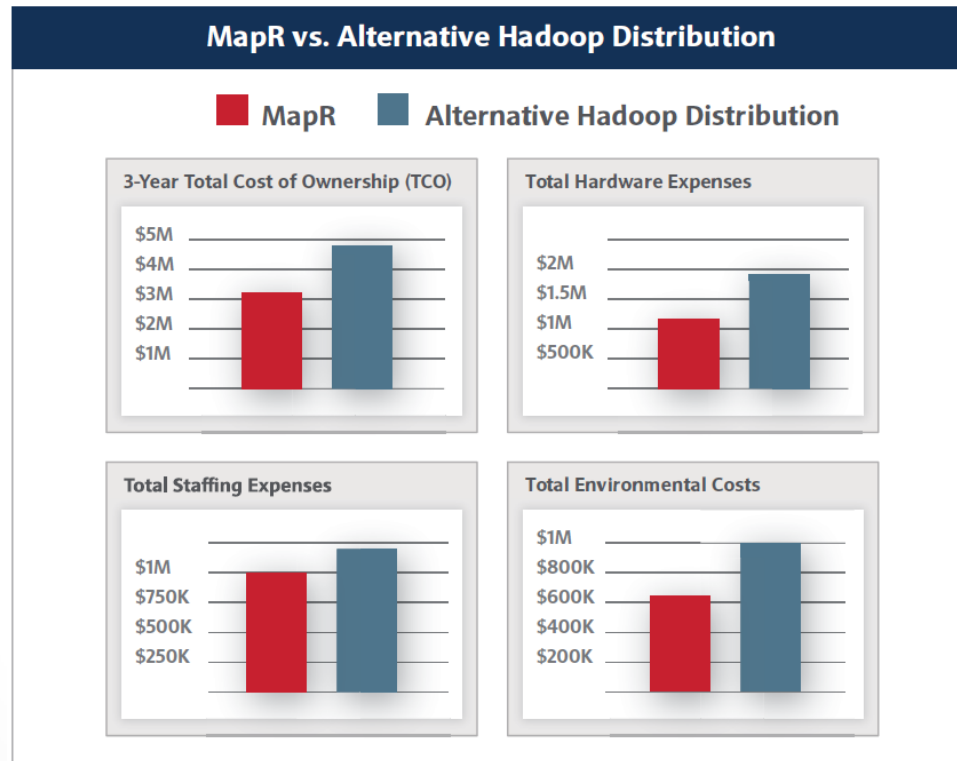


Figure 7. Cost comparison for a 500 TB cluster between two commercial Hadoop distributions. Copyright 2014 by MapR. Reprinted with permission

The IaaS option

Infrastructure as a Service (IaaS)

One consideration that can mitigate the cost implications of hardware and support personnel is the use of a cloud offering.

As pointed out by Intel (2015) clouds are already deployed on pools of server, storage, and networking resources and can scale up or down as needed. Cloud computing offers a cost-effective way to support Big Data technologies and the advanced analytics applications that can drive business value.

Diversity Limited (2010) defines Infrastructure as a Service (IaaS) as “a way of delivering Cloud Computing infrastructure – servers, storage, network and operating systems – as an on-demand service. Rather than purchasing servers, software, datacenter space or network equipment, organisations instead buy those resources as a fully outsourced service on demand.”

Selecting a Hadoop distribution

According to CITO Research (2014):

Core distribution: All vendors use the Apache Hadoop core and package it for enterprise use.

Management capabilities: Some vendors provide an additional layer of management software that helps administrators conjure, monitor, and tune Hadoop.

Enterprise reliability and integration: A third class of vendors offers a more robust package, including a management layer augmented with connectors to existing enterprise systems and engineered to provide the same high level of availability, scalability, and reliability as other enterprise systems.

Selecting a Hadoop distribution

According to Ovum (2014), as an emerging platform, commercial distributions are differentiating; organizations should closely scrutinize their platform supplier's strategy for delivering an enterprise-grade platform. If Hadoop is to become a platform on which enterprises store system of record data, and rely on for competitive analytics insights and operational applications or decisions, it must deliver:

- Predictable performance -- Deliver consistent performance/availability/reliability to meet business SLAs.
- Security – Provides the same degree of granular security addressing access rights, privacy, and protection unauthorized actions as enterprise databases.
- Data protection – Ensure the same degree (e.g., backup and recovery, privacy, access, obfuscation, and auditing of usage) that is offered by enterprise databases.
- Data governance and stewardship -- Support the policy-driven lifecycle management of data, like any enterprise database.

(Ovum, 2014, pg. 2)

Selecting a Hadoop distribution

Zikopoulos et al. (2013) believe that any Big Data platform must include the six key imperatives that are shown in Figure 6.






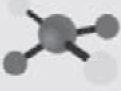
	Big Data Platform Imperatives		Technology Capability
1	Discover, explore, and navigate Big Data sources		Federated Discovery, Search, and Navigation
2	Extreme performance—run analytics closer to data		Massively Parallel Processing Analytic appliances
3	Manage and analyze unstructured data		Hadoop File System/MapReduce Text Analytics
4	Analyze data in motion		Stream Computing
5	Rich library of analytical functions and tools		In-Database Analytics Libraries Big Data Visualization
6	Integrate and govern all data sources		Integration, Data Quality, Security, Lifecycle Management, MDM, etc

Figure 8. Imperatives and underlying technologies. Copyright 2014 by McGraw-Hill Education. Reprinted with permission

There is a lot out there.....

451 Research

Data Platforms Map June 2015

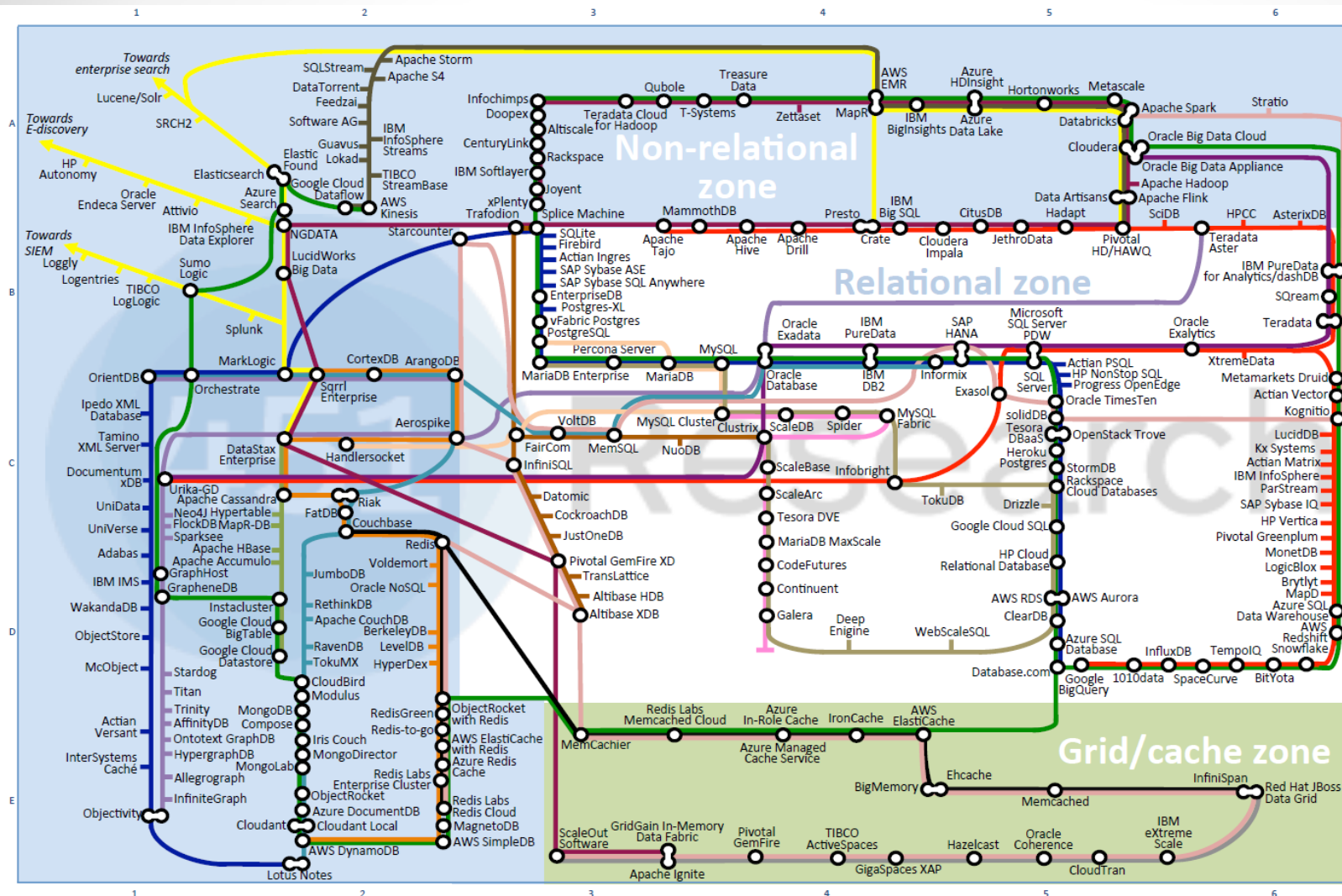
Key:

- General purpose
- Specialist analytic
- as-a-Service
- BigTables
- Graph
- Document
- Key value stores
- Key value direct access
- Hadoop
- MySQL ecosystem
- Advanced clustering/sharding
- New SQL databases
- Data caching
- Data grid
- Search
- Appliances
- In-memory
- Stream processing

<https://451research.com/dashboard/dpa>

451research.com/
dashboard/dpa

© 2015 by 451 Research LLC.
All rights reserved



Hadoop distributions

Forrester (2014) lists the following general-purpose big data Hadoop distributions:

Vendor	Product
Amazon Web Service	Amazon Elastic MapReduce
Cloudera	Cloudera Enterprise
Hortonworks	Hortonworks Data Platform
IBM	InfoSphere BigInsights
Intel	Intel Distribution for Apache Hadoop
MapR Technologies	MapR M3 - MapR M7
Microsoft	Windows Azure HDInsight
Pivotal Software	Pivotal HD
Teradata	Teradata Open Distribution for Hadoop (TDH)

Selecting a Hadoop distribution

Shortlist:

- Hortonworks
- Cloudera
- MapR
- IBM

Final selection:

- MapR
- IBM

Recommendation:

- IBM

CITO Research: Hortonworks

According to CITO Research (2014) some distributions predicate their value proposition on their proximity to the original Hadoop trunk, the core distribution of the software committed to a source-code management system. Hortonworks, founded by Yahoo! Engineers who had worked on the original implementation of Hadoop, offers a services-only model for Hadoop. Its objective is to stay as close to the original open-source trunk of Hadoop as possible, limiting any enhancements to its product to those offered publicly by the Hadoop community.

Hortonworks charges for support, but not for the distribution itself. Their rationale is twofold. First, Hortonworks believes that the Hadoop community, to which Hortonworks is a contributor, will devise the optimum solution. (It's worth noting that in practice the community has not changed Hadoop much in the past two years.) Hortonworks also believes that it's important to avoid vendor lock-in by committing to a forked version of Hadoop.

Hortonworks is focused on improving the usability of the Hadoop platform as it is written and presented in the Apache repository by offering a free distribution and subscription support. This approach works for parties with a vested interest in Hadoop remaining open-source but also reliable.

The services model seems to be geared toward developers committed to using vanilla Hadoop as their data management platform. However, the Apache distribution is not optimized to support snapshots (point-in-time images of the entire file system or sub-trees thereof, often used for error recovery), Network File System (NFS) integration, or real-time data streaming into the cluster. Such enhancements require re-architecting the underlying infrastructure. (CITO Research, 2014, pg. 6)

CITO Research: Cloudera

According to CITO Research (2014) distributions with added management capabilities, typified by Cloudera, seek to improve on one area of Hadoop while leaving the rest to the open-source community. Cloudera adds capabilities to Hadoop based on small variations in the trunk. Cloudera offers a subscription service that packages Hadoop and supplies support and management software that helps administrators configure, monitor, and tune Hadoop. The company also offers training and consulting services that fill the gap between what the community can provide and what enterprises need to integrate Hadoop as part of their data management strategy.

Cloudera's focus on the management layer, to the exclusion of all other aspects of Hadoop that prevent it from being truly enterprise-ready, seems a limited proposition. Cloudera counts on the open-source community—including contributors from Cloudera, Hortonworks, LinkedIn, Huawei, IBM, Facebook and Yahoo!—to innovate faster than vendors that change the distribution in a significant way. Cloudera's leadership also asserts that the strong Apache governance model can resolve many of the conflicts in Hadoop's development path that may arise. However, recent history raises some questions since the flaws in Hadoop's architecture have not been addressed over the past 7 years.

(CITO Research, 2014, pg. 7)

Gartner



Figure 10. Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics . Copyright 2015 by Gartner. Reprinted with permission

Gartner: IBM Strengths

IBM demonstrates a broad offering and integration across products that can support all four major data warehouse use cases. In addition, IBM's approach to the cloud (which includes a solution for data integration) and transformation in the cloud sets a new tone for analytics in the market. IBM has continued to invest in product innovation and meeting emerging customer and market demands such as the delivery of its data warehouse PaaS offering dashDB, including in-memory columnar capabilities and integrated with its Cloudant NoSQL database (also delivered as a cloud service), as well as PureData for in-database analytics. Customer references report higher levels of satisfaction than in the past regarding pricing and value for money, which demonstrates IBM's intention to compete more aggressively on price.

(Gartner, 2015, pg. 4)

Gartner: IBM Cautions

Overall, IBM suffers from complex marketing and branding that confuses the market. For example, the rebranding of products under Watson Foundations and "cognitive" causes some confusion because it is unclear what value Watson brings relative to customers' demands.

By the end of 2013, IBM's worldwide market share ranking for database software dropped to No. 3. While the data warehouse market is extending to include DMSA, the relational DBMS portion of the market constitutes over 90% of the combined market. During 2014, Gartner inquiry clients expressed unease regarding IBM's commitment to traditional database delivery. Organizations need to carefully decide, when taking a best-fit engineering approach, what combination of technology aligns with IBM's offering.

IBM customers are unclear how to move from traditional IBM offerings — that combine software licenses, appliances (PureData, formerly Netezza) and on-premises concepts — to modern infrastructure deployment (such as "data refinery"). IBM has been driving innovation, addressing new demands for data preparation, preserving governance and differentiating from data-lake approaches; however, it remains unclear how existing data warehouse DBMS investments and new information approaches live side by side, or how to correctly identify the appropriate solutions for specific use cases.

(Gartner, 2015, pg. 5)

Gartner: MapR Strengths

Based on information provided by MapR, it is the Hadoop distribution vendor with the largest number of paying customers — which is an important indication of adoption for an emerging category of technology.

MapR's strategy is to deliver a data platform that combines Hadoop and operational database technologies to support a wide range of workloads in a single deployment. To enable this strategy, the company has compensated for Hadoop deficiencies by creating alternatives to Apache components while including a number of open-source projects from other distributions. For example, it substitutes Hadoop Distributed File System (HDFS) with its Posix-compliant, standard Network File System (NFS) file system, and it also supports Impala. This inclusive strategy offers customers the greatest number of options.

MapR is praised by references for its reliability, performance and scalability, making it a solution suitable for enterprise use.

(Gartner, 2015, pg. 6)

Gartner: MapR Cautions

MapR has a smaller partner ecosystem than the other Hadoop distribution vendors. For example, the number of DBMS, BI or data integration partners is modest. However, MapR is actively addressing this by recently adding partnerships with Teradata, SAS and HP Vertica (for example).

Reference customers struggle to find enough skilled resources in the market. The growing interest in Hadoop will help to relieve some of this pressure, but this is a multiyear cycle rather than one measured in quarters or months.

Reference customers indicate that it can take time for MapR to support the latest Hadoop capabilities, although it can support multiple versions of the same Hadoop project. To address this concern, MapR accelerated its ecosystem update process during March 2013, and now has monthly Hadoop ecosystem releases.

(Gartner, 2015, pg. 6)

Forrester Wave

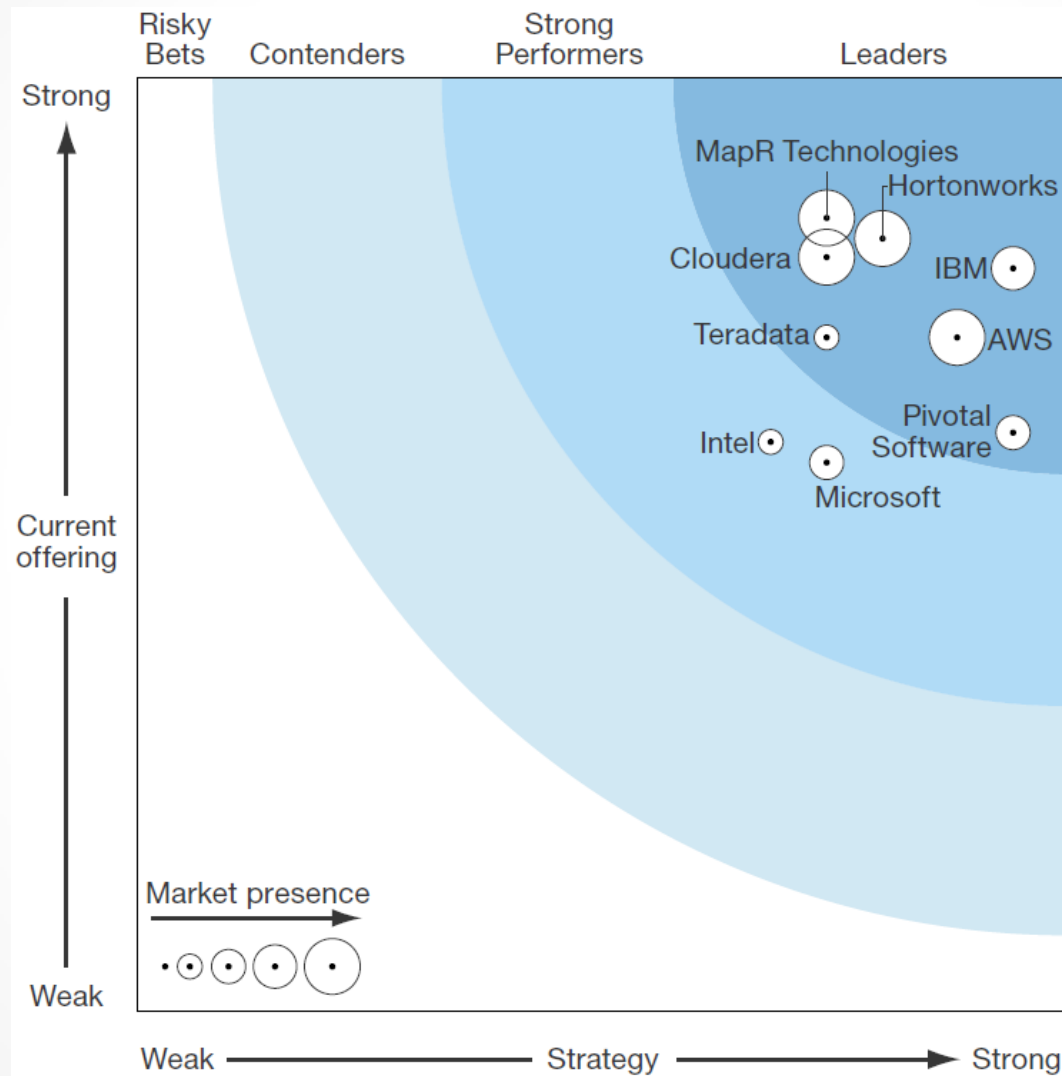


Figure 11. Forrester Wave™: Big Data Hadoop Solutions, Q1 '14. Copyright 2014 by Forrester Research Inc. Reprinted with permission

Forrester Wave: IBM

IBM flexes its enterprise muscles with InfoSphere BigInsights.

Distributed computing platforms and data management are certainly not new to IBM. It has offerings in grid computing, databases, and many other data management technologies that it can bring to a comprehensive Hadoop solution. In addition, IBM has advanced analytics tools, a global presence, and implementation services, so it can offer a complete big data solution that will be attractive to many customers. IBM's road map includes continuing to integrate the BigInsights Hadoop solution with related IBM assets like SPSS advanced analytics, workload management for highperformance computing, BI tools, and data management and modeling tools. Today, IBM has more than 100 Hadoop deployments, some of which are fairly large and run to petabytes of data.

(Forrester Research Inc, 2014, pg. 10)

Forrester Wave: MapR

MapR Technologies scored highest for its current offering of all the vendors.

The score speaks for itself. MapR Technologies has added some unique innovations to its Hadoop distribution, including support for Network File System (NFS), running arbitrary code in the cluster, performance enhancements for HBase, as well as high-availability and disaster recovery features. However, MapR Technologies has lagged behind the other pure-play vendors, Cloudera and Hortonworks, in terms of market awareness; Forrester clients often ask about Cloudera and Hortonworks — but not about MapR Technologies. MapR Technologies has a leading solution; it must now make more noise in the market and accelerate its partnerships and distribution channels.

(Forrester Research Inc, 2014, pg. 10)

CITO Research: MapR

Enterprise Reliability and Integration – MapR Technologies

The enterprise reliability and integration category, typified by MapR Technologies, is built so that Hadoop connects to as many processing sources and consuming applications as possible. It also gives Hadoop the same high availability, scalability, and reliability as other enterprise systems. Hadoop lets organizations process data at scale, but requires other platforms to interpret and extract value from that data. The founders of MapR asserted that Hadoop was too important to leave it in an underdeveloped state. While acknowledging the power of open-source development, MapR operates under the belief that a market-driven entity is more likely to support market needs, faster.

While maintaining the core distribution, MapR has also conducted proprietary development in some critical areas where the open-source community has not solved Hadoop's flaws.

(CITO Research, 2014, pg. 7)

CITO Research: MapR

Improving HDFS for high performance and high availability.

HDFS does not support enterprise-grade performance, and MapR sought to change that in a number of ways.

- MapR replaced HDFS so that it would not be reliant on Java or on the underlying Linux file system.
- MapR's version of HDFS allows dynamic reading and writing, whereas the original HDFS is an append-only file system that can only be written once.
- MapR de-centralized and distributed the NameNode, increasing its capacity from 100 million files to 1 trillion. Because each node contains a copy of the NameNode, all nodes can participate in failure recovery, instead of requiring each node to call back to a central instance.

The choice of distribution depends largely on what represents an obstacle to a given organization's plans for Hadoop. For the organization that is interested in participating in the continued development of Hadoop, choosing a distribution that makes limited or no alterations to the trunk is a logical vote. For the organization that needs to connect Hadoop to multiple data analysis platforms, has a requirement for reliability, visibility, and control for mission-critical applications, a comprehensive solution that is available from a vendor in the here and now might make more sense.

(CITO Research, 2014, pg. 11)

MapR Positioning

- Performance-optimized architecture for faster data processing and analytics
- Direct Access NFS™ for real-time data access to Hadoop data
- Distributed metadata to support one trillion files in a single cluster
- MapR Heatmap™ for instant cluster insights
- MapR volumes for easier policy management around security, retention, and quotas

(MapR, 2015, pg. 1)

IBM BigInsights Overview

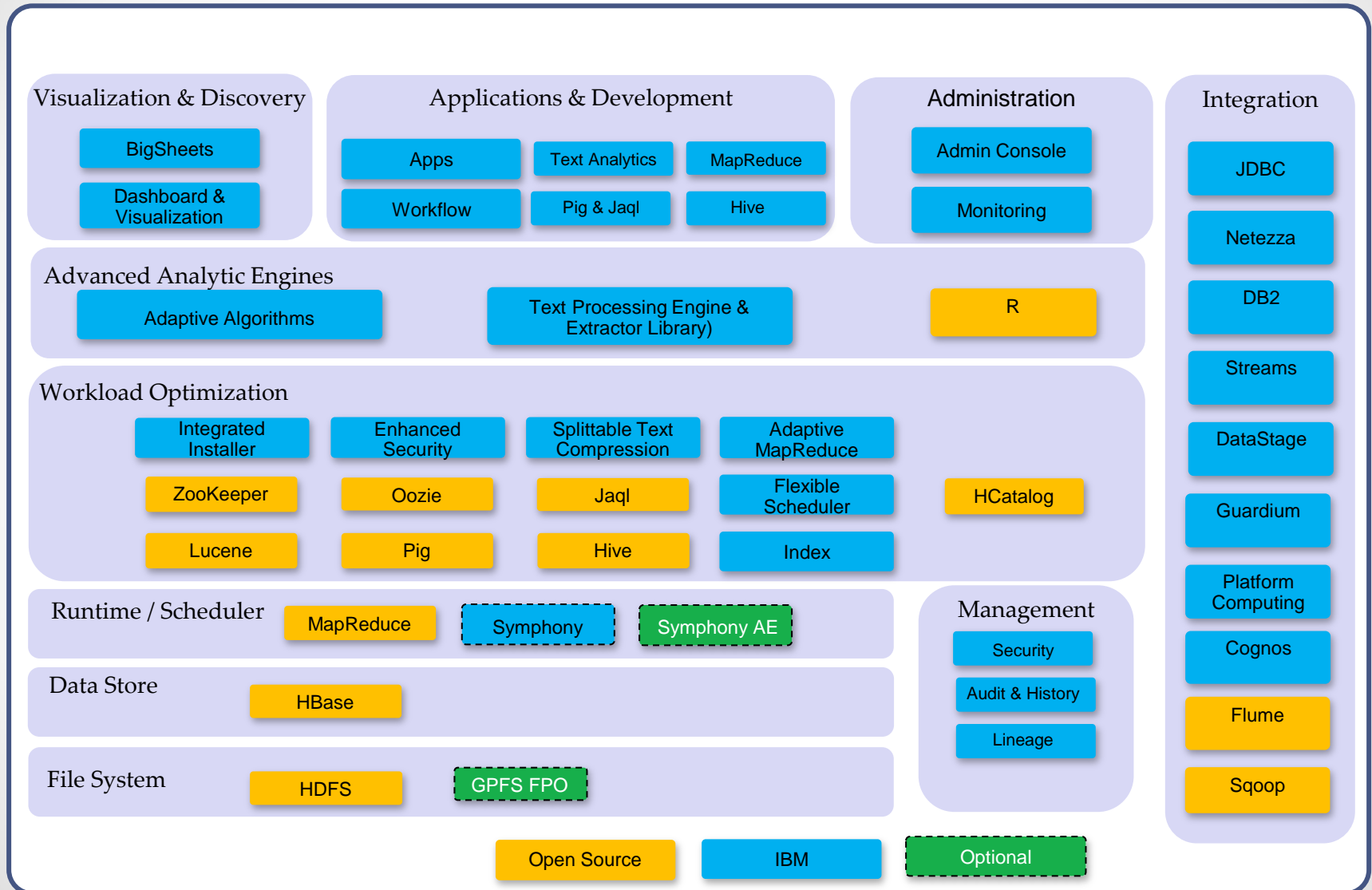


Figure 12. IBM InfoSphere BigInsights. Copyright 2013 by IBM. Reprinted with permission

IBM BigInsights for Apache Hadoop Offering Suite

IBM BigInsights v4	BigInsights Quick Start Edition	IBM Open Platform with Apache Hadoop	Elite Support for IBM Open Platform with Apache Hadoop	BigInsights Analyst Module	BigInsights Data Scientist Module	BigInsights Enterprise Management Module	BigInsights for Apache Hadoop
Apache Hadoop Stack: HDFS, YARN, MapReduce, Ambari, Hbase, Hive, Oozie, Parquet, Parquet Format, Pig, Snappy, Solr, Spark, Sqoop, Zookeeper, Open JDK, Knox, Slider	✓	✓	✓	*	*	*	✓
				* Paid support for IBM Open Platform with Apache Hadoop required for BigInsights modules			
Big SQL – 100% ANSI compliant, high performant, secure SQL engine	✓			✓	✓		✓
BigSheets – spreadsheet-like interface for discovery & visualization	✓			✓	✓		✓
Big R – advanced statistical & data mining	✓				✓		✓
Machine Learning with Big R – machine learning algorithms apply to Hadoop data set	✓				✓		✓
Advanced Text Analytics – visual tooling to annotate automated text extraction	✓				✓		✓
Enterprise Mgmt – Enhanced cluster & resource mgmt & POSIX-compliant file systems						✓	✓
Governance Catalog				✓	✓		✓
Cognos BI, InfoSphere Streams, Watson Explorer, Data Click							✓

Figure 13. IBM BigInsights for Apache Hadoop Offering Suite. Copyright 2014 by IBM. Reprinted with permission

BigInsights

IBM BigInsights is based on 100 percent open source Hadoop. It extends Hadoop with enterprise-grade technology including administration and integration capabilities, visualization and discovery tools as well as security, audit history and performance management.

According to IBM, the BigInsights platform offers:

- Increased performance: Out-performing open source alternatives on identical hardware, IBM InfoSphere BigInsights has been independently benchmarked and proven to be between 4 and 11 times faster than open source alternatives running on identical infrastructure. InfoSphere BigInsights provides several features that help increase performance, as well as enhance its adaptability and compatibility within an enterprise environment. (IBM, 2014)
- Usability: BigInsights is optimized for a wide range of roles, including integration developers, administrators, data scientists, analysts and line-of-business contacts.
- Integrated with IBM Watson™ Foundations big data platform: BigInsights comes bundled with search and streaming analytics capabilities.
- Analytics: Built-in Hadoop analytics capabilities for machine data, social data, text and Big R enable you to locate actionable insights from data in the Hadoop cluster rather than having to move the data around.

BigInsights components

IBM BigInsights. Hadoop offering that provides open source Hadoop components along with additional enterprise-class capabilities:

- InfoSphere Streams. World's first, industry leading, ultra low latency streaming analytics platform
- BigSQL. A full-featured, high-performance SQL engine for Hadoop
- Text Analytics Engine. Highly accurate framework and engine for unstructured data analysis
- BigR. R deployed on Hadoop to leverage Hadoop distributed processing capabilities
- Big Match. Highly scalable entity matching for master data management
- Identity Insights (G2). Context computing engine
- IBM Research innovations in the Early Access Program. Machine learning, graph analytics, sentiment analysis

(IBM, 2014)

BigInsights integration

Integrated business functionality is delivered through the breadth of the IBM Big Data Platform which includes the following capabilities:

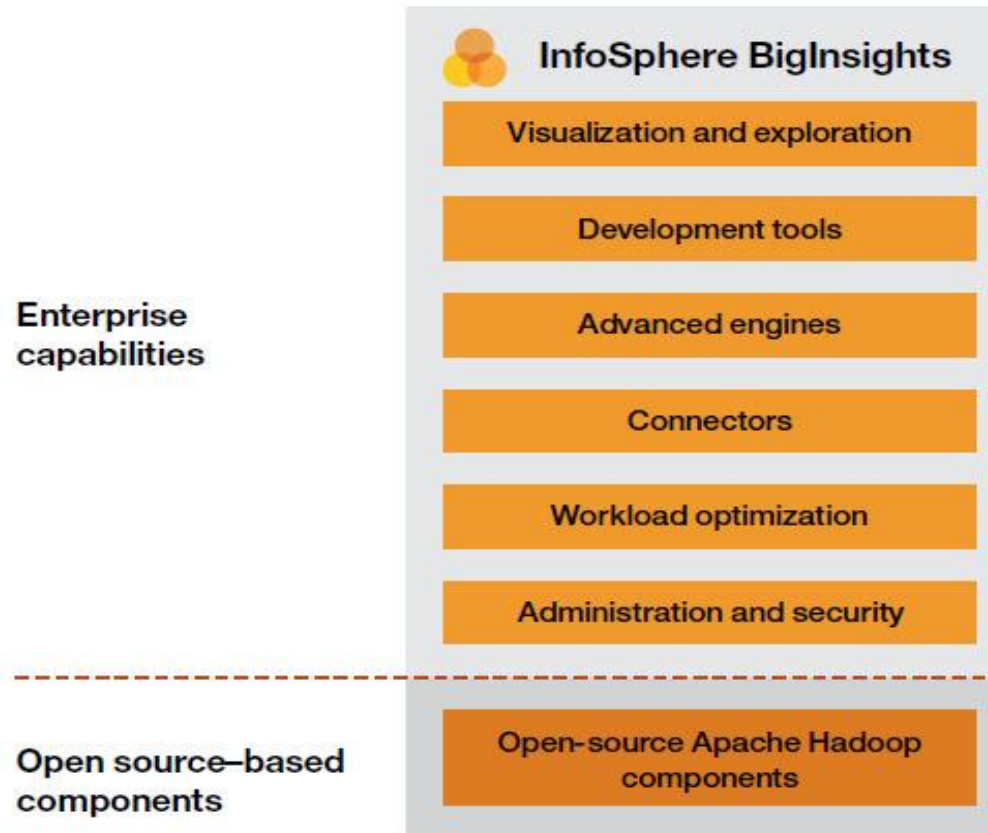


Figure 14. Big Data Platform integration. Copyright 2014 by IBM. Reprinted with permission

A note on open source

Fully Open Source or Only Skin Deep?

There are two vitally important elements for open source software: licensing and development. The Apache license is a commercially (and user) friendly licensing model, which means anyone can use or even modify the code as long as they attribute where it comes from. The ASF has a rigorous set of procedures and standards ensuring that Apache software projects are run in an open and democratic manner for the benefit of the project's community. The idea is to prevent individuals or corporate agendas from taking over a project. So when people describe a software offering as being open source, investigate whether it's just the licensing or the development model as well. For example, Cloudera Impala is often touted as an open source project but is open source only from a licensing perspective. Cloudera engineers are the only people contributing code to this project. An open source license is well and good, but from a business perspective, that's not where the real benefits of open source code lie. (This isn't to besmirch Cloudera Impala; we think the IBM Big SQL technology does that on its own. Rather, it's to give you a framework to understand when something is open source and when something isn't. Both IBM and Cloudera have tremendous contributions to open source as well as their own proprietary technology.)

Apache Hadoop, for example, has contributions from hundreds of developers from dozens of different companies, all working together in a public manner. This has resulted in rapid innovation and the satisfaction of requirements from many different corners of the community. That's the real value of open source for businesses.

(Zikopoulos, deRoos, Bienko, Buglio, Andrews, 2014, pg. 138)

IBM and open source

According to IBM (2014):

We distribute 100% open source Apache Hadoop components. This is not proprietary. On top of the open source code we provide analytical tools to help get value from the data.

IBM is committed to supporting the open source movement. IBM helped open platforms such as Linux, Eclipse and Apache become standards with vital industry ecosystems, and then we developed high-value businesses on top of them. Today IBM collaborates broadly to support open platforms such as OpenStack and Hadoop.

Because of this commitment, IBM avoids creating any independent fork of Apache project code, and merely selects the open source versions that we feel are the best in achieving most current and most stable capabilities together in the overall Hadoop operating environment. The inner core of BigInsights is Apache Hadoop, and we do inter-version testing of the projects included so our enterprise customers are ensured that they have a blue-washed and interoperable codebase across the projects. As “most current” and “stable” are often conflicting, BigInsights does not always use the most current version of projects, but rather the most stable. Where we identify issues in the open source projects, we have a number of committers with our IBM development labs that submit fixes back to the open source community.

IBM's goal with this approach is to protect the corporate IT organisation from version management across the various open source projects by providing this pre-tested, interoperable set in InfoSphere BigInsights. An example of the commitment is that IBM contributed 25% the fixes for a recent release of Hadoop.

(IBM, 2014, pg. 13)

BigInsights vs Hortonworks

According to IBM (2014) BigInsights and Hortonworks have similar Hadoop components and both are committed to open source Apache Hadoop with Committers and contributors to open source Apache Hadoop. However, BigInsights extends value beyond Hortonworks for analytics with its Social Media Accelerator, Machine Data Accelerator, BigSheets spreadsheet and visualization, Advanced Text Analytics. Also BigInsights includes Data Explorer for Search and Indexing in Hadoop and beyond to all enterprise data; a vital function to make the data accessible to potential users inside the company and through applications to its customers. BigInsights Big SQL has advantages over Horton's HiveQL as Big SQL provides richer SQL, HBase performance, and short query performance.

For many large companies already using GPFS, IBM BigInsights 2.1 uniquely offers GPFS as a Hadoop file system providing enterprise data life cycle management. BigInsights also has Adaptive MapReduce (Platform Symphony) for faster Map Reduce processing and BigInsights integrates with InfoSphere Streams, while Hortonworks does not, limiting its use to batch processing.

(IBM, 2014, pg. 16)

BigInsights vs Cloudera

According to IBM (2014), IBM has a comparable number of significant sized deployments with Cloudera, a Hadoop distributor. However the company is quite different to IBM. Cloudera is Venture Capital funded with \$160m invested in its 6th funding investment round completed in March 2014. Sales revenue was reported as \$73m in 2013, its fourth year trading. It has 500 employees.

Our opinion on the long term destiny for companies like this – niche technology players – is of a business exit plan based on acquisition by the industry giants to fill a technology gap in the enterprise platforms they provide. At this point it's not clear if any of the enterprise technology mega vendors have such a gap so the future of Cloudera is unclear.

Functionality which will be useful to the vehicle data industry such as real time streaming data analytics, text analytics, analytics accelerator tools, visualisation, enterprise wide search, indexing, data integration software, connected analytics appliances, relational data marts, governance audit and compliance is all available in IBM BigInsights but not in this alternative.

Documented limitations in the Cloudera query engine explain why results in data joins can fail to complete. Referring to the Cloudera user manual highlights the cause as insufficient memory – this is caused by the need to load ALL data into memory. As raw data sets can be very large this limitation can easily exceed the total memory available. Vehicle data volumes being generated are currently very large, and customer datasets are also very large so this limitation constrains vehicle industry applications. By comparison IBM BigSQL has no limitation that joined tables have to fit in aggregated memory of data nodes which causes queries to run out of memory and fail. IBM Hadoop is up to 41x faster than Hive .12 (Cloudera) on TPC-H like benchmark IBM Hadoop is over 2x faster than (Cloudera) Impala on TPC-H like Benchmark

(IBM, 2014, pg. 15)

BigInsights vs Pivotal /EMC/Greenplum

Greenplum has changed hands several times and is now part of Pivotal, an EMC spinoff, and their Hadoop offering is now called Pivotal HD. IBM BigInsights has many advantages over Pivotal HD

IBM BigInsights adds significant functions beyond IBM's 100% open source Hadoop components – it includes analytic accelerators such as Big SQL, BigSheets, BigMatch, BigR and text analytics unlike Pivotal which includes proprietary components and lack of added-value software applications such as those listed above IBM has already achieved broader marketplace presence and analyst rating (e.g. Forrester Wave) IBM BigInsights offers greater flexibility and lower cost solution with availability as software only, on the cloud, or on flexible IBM System x reference architecture. By comparison Pivotal is now recommending expensive Isilon storage which uses a proprietary OneFS file system. IBM has made significant investments to ensure enterprise open architecture leverage the low cost Hadoop elements rather than creating lock-in solutions.

Significantly, Pivotal does not support HDFS. BigInsights offers HDFS and GPFS support. Where the new SQL HAWQ component of Pivotal HD is offered as a license cost option, the powerful IBM Big SQL is included with BigInsights. IBM offers a complete Big Data platform Solution as an integrated architecture that offers more than just Hadoop – including BigInsights, Streams, MPP Database, Information Integration. Real time analytics – not just batch – is provided by IBM, whereas Pivotal has an in-memory grid, which is not a real time streaming solution.

Data security at an enterprise granular level is provided by IBM's Integration and Governance offerings, (Information Server, Guardium and Optim) which are integrated with Hadoop but would require local development or 3rd party integration projects to achieve the same level of data management.

As the vehicle data will include personal data, its governance is mandated. Delivering the appropriate systems is easier and lower cost with IBM. Pivotal HAWQ adds the entire RDBMS structure (query engine, storage layer, metadata) to Hadoop. This adds proprietary layers and database complexity to the Hadoop solution. By comparison IBM Big SQL integrates just the query engine with Hadoop. This allows the query engine to be collocated with the Hadoop cluster and executes using native meta data and HDFS files, which is how IBM won the performance benchmark tests cited above, and IBM Big SQL offers elastic scalability where nodes can be added / removed online.

Pricing and licensing

Products	BigInsights Quick Start	IBM Open Platform with Apache Hadoop	Elite Support for IBM Open Platform with Apache Hadoop	BigInsights Analyst Module	BigInsights Data Scientist Module	BigInsights Enterprise Management Module	BlgInsights for Apache Hadoop
Pricing Terms	Free	Free	Yearly Subscription Only	Perpetual or Monthly License			
Support provided	Community	Community	IBM 24x7 support				
Usage License	Non-production, five node cap	Production Usage					
Pricing Model	Free	Free	Node based pricing				
Access via	ibm.com/hadoop		Passport Advantage				

Figure 15. Pricing and licensing. Copyright 2014 by IBM. Reprinted with permission

ITG: IBM InfoSphere BigInsights Cost

Three-year Costs for Use of IBM InfoSphere BigInsights and Open Source Apache Hadoop for Major Applications – Averages for All Installations

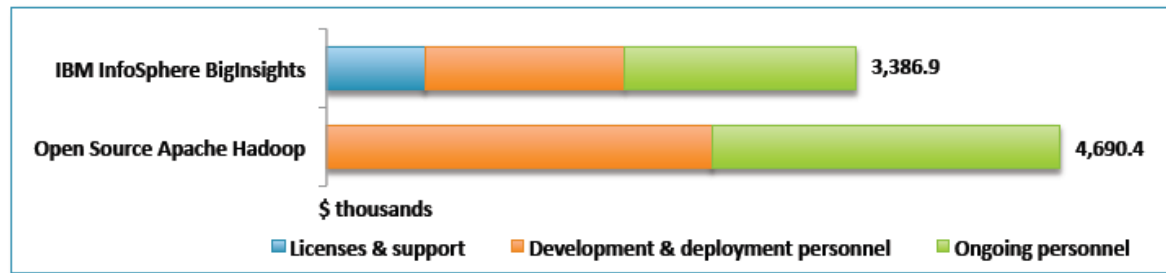
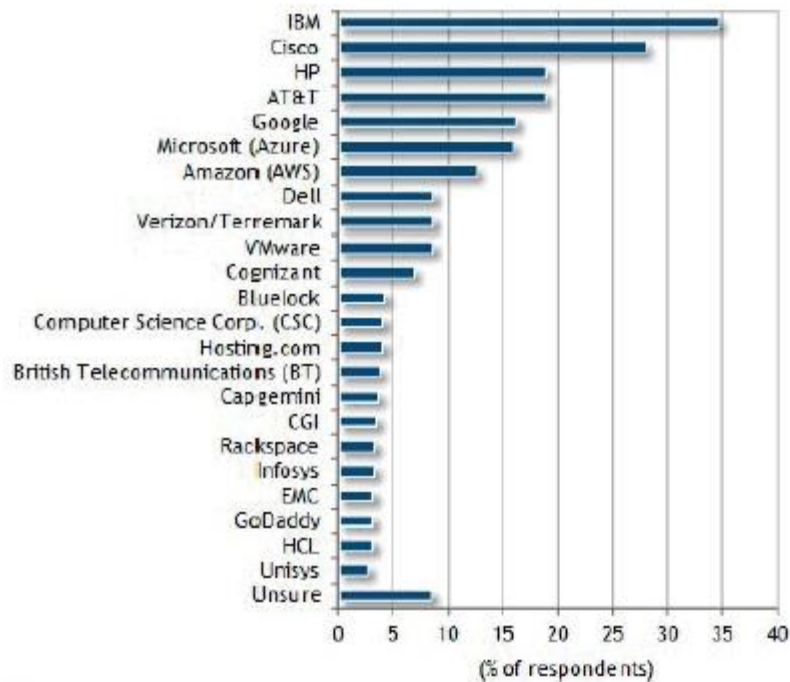


Figure 16. Three-year Costs for Use of IBM InfoSphere BigInsights and Open Source Apache Hadoop for Major Applications. Copyright 2013 by the International Technology Group. Reprinted with permission.

An IaaS provider

U.S. Top IaaS Provider Preferences

Q. Please select the top 3 vendors that you believe can provide infrastructure as a service (IaaS), for private and/or public, most effectively, whether or not you currently utilize these services from third-party providers.



n = 402

Figure 17. U.S. Top IaaS Provider Preferences. Copyright 2013 by IDC. Reprinted with permission

Summary

Big Data is having a substantive impact on the P&C insurance industry. Insurers are combining Big Data and analytics to overcome many of the challenges confronting the industry, and to support new capabilities. Although implementing a Big Data platform is not without its' challenges, the availability of IaaS platforms for Big Data reduce many of the initial risks that would traditionally be associated with such projects.

IBM is a leading IaaS provider and together with the IBM BigInsights platform provide a compelling Big Data offering. While the competition around Hadoop distributions is strong, particularly from MapR Technologies, it is recommended that if only one Hadoop distribution is to be taken forward for proof-of-concept, that it be the IBM BigInsights platform. Should two solutions be considered for POC, then MapR technologies should be added to the evaluation.

References

- CITO Research. (2014). Choosing a provider from the hadoop ecosystem. [pdf]. Retrieved from <http://www.citoresearch.com>
- Diversity Limited. (2010). Moving your infrastructure to the cloud. [pdf]. Retrieved from <http://diversity.net.nz/wp-content/uploads/2011/01/Moving-to-the-Clouds.pdf>
- Forrester Research Inc. (2014). *Forrester wave: big data hadoop solutions, Q1 '14*. [diagram]. Retrieved from Forrester Research Inc. (2014). *Forrester wave: big data hadoop solutions, Q1 '14* [pdf]. Retrieved from <http://www.forrester.com>
- Gartner. (2015). Magic quadrant for data warehouse and data management solution. [pdf]. Retrieved from <http://www.gartner.com>
- Gartner. (2015). *Magic quadrant for data warehouse and data management solution*. [diagram]. Retrieved from Gartner. (2015). *Magic quadrant for data warehouse and data management solution*. [pdf]. Retrieved from <http://www.gartner.com>
- IBM. (2013). *Big difference: schema on run*. [diagram]. Retrieved from IBM. (2013). Overview - big data & analytics[pdf]. Retrieved from <http://www.slideshare.net/vmanoria1/overview-ibm-big-data-platform>
- IBM. (2013). *IBM infoSphere bigInsights*. [diagram]. Retrieved from IBM.. (2013). Best practices for deploying infosphere bigInsights and infoSphere streams in the cloud. [ppt]. Retrieved from <http://www.slideshare.net/leonsp/best-practices-for-deploying-hadoop-biginsights-in-the-cloud>
- IBM. (2014). IBM expands hadoop commitment with support for spark. [blog]. Retrieved from <http://www.ibmbigdatahub.com/blog/ibm-expands-hadoop-commitment-support-spark>
- IBM. (2014). *IBM BigInsights for apache hadoop offering suite*. [diagram]. Retrieved from IBM. (2014). IBM's bigInsights: smart analytics for big data. [ppt]. Retrieved from <http://www.slideshare.net/CynthiaSaracco/introducing-ibms-info>
- IBM. (2014). *IBM pricing and licensing*. [diagram]. Retrieved from IBM. (2014). IBM's bigInsights: smart analytics for big data. [ppt]. Retrieved from <http://www.slideshare.net/CynthiaSaracco/introducing-ibms-info>
- IBM. (2014). The top 10 reasons manufacturers should use BigInsights as major vehicle IBM their Hadoop Platform. [pdf]. Retrieved from <http://www.slideshare.net/TobyWoolfe/2014-10-09-top-reasons-to-use-ibm-biginsights-as-your-big-data-hadoop-system-42638020>
- IBM. (2014). *Big data platform integration*. [diagram]. Retrieved from IBM. (2014). The top 10 reasons manufacturers should use BigInsights as major vehicle IBM their Hadoop Platform. [pdf]. Retrieved from <http://www.slideshare.net/TobyWoolfe/2014-10-09-top-reasons-to-use-ibm-biginsights-as-your-big-data-hadoop-system-42638020>
- IBM. (2015). *Untitled*. [Diagram]. Retrieved from IBM. (2015). Unlocking business value through enterprise hadoop adoption [pdf]. Retrieved from <http://www.ibmbigdatahub.com/blog/unlocking-business-value-through-enterprise-hadoop-adoption>

References

- IBM. (2015). Analytics: what is mapreduce. [web page]. Retrieved from <http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/>
- IDC. (2013). *U.S. top iaas provider preferences*. [response chart]. Retrieved from IBM. [2014]. Report: IBM named #1 preferred provider report: IBM named #1 preferred provider of iaas cloud by enterprises. [news release]. Retrieved from <https://www-03.ibm.com/press/us/en/pressrelease/43893.wss>
- Intel. (2015). Big data cloud technology. [pdf]. Retrieved from <http://www.intel.co.za/content/dam/www/public/us/en/documents/product-briefs/big-data- cloud-technologies-brief.pdf>
- ITG. (2013). Three-year costs for use of IBM infosphere bigInsights and open source apache hadoop for major applications. [Diagram]. Retrieved from ITG. (2013). Business case for enterprise big data deployments. [pdf]. Retrieved from <http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IME14028USEN&appname=skmwww>
- ITG. (2013). Business case for enterprise big data deployments. [pdf]. Retrieved from <http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IME14028USEN&appname=skmwww>
- MapR. (2014). *Calculate your total cost of ownership of apache hadoop*. [diagram]. Retrieved from MapR. (2014). Calculate your total cost of ownership of apache hadoop. [pdf]. Retrieved from https://www.mapr.com/sites/default/files/solutionsbrief_tco.pdf
- MapR. (2014). Calculate your total cost of ownership of apache hadoop. [pdf]. Retrieved from https://www.mapr.com/sites/default/files/solutionsbrief_tco.pdf
- MapR. (2015). The mapr distribution including apache hadoop community edition. [pdf]. Retrieved from <https://www.mapr.com>
- Ovum. (2014). Enterprise-grade hadoop. [pdf]. Retrieved from www.ovum.com
- Planet Cassandra. (2015). Nosql databases defined and explained. [web page]. Retrieved from <http://www.planetcassandra.org/what-is-nosql/>
- TDWI Research. (2015). Hadoop for the enterprise. [pdf]. Retrieved from <http://www-01.ibm.com/software/data/infosphere/hadoop/resources.html>
- Zikopoulos, P., deRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., Giles, J. (2013). Harness the power of big data. [pdf]. Retrieved from <http://www-01.ibm.com/software/de/big-data/pdf/assets/Harness.PDF>
- Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., Andrews, M. (2014). Big data beyond the hype. [pdf]. Retrieved from https://www.ibm.com/developerworks/community/blogs/SusanVisser/entry/big_data_beyond_the_hype_a_guide_to_conversations_for_today_s_data_center?lang=en

References

- Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., Andrews, M. (2014)). *A next-generation zones reference architecture for big data and analytics*. [diagram]. Retrieved from Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., Andrews, M. (2014). *Big data beyond the hype*. [pdf]. Retrieved from https://www.ibm.com/developerworks/community/blogs/SusanVisser/entry/big_data_beyond_the_hype_a_guide_to_conversations_for_today_s_data_center?lang=en
- Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., Andrews, M. (2014)). *Characteristics of the nosql and sql worlds*. [table]. Retrieved from Zikopoulos, P., deRoos, D., Bienko, C., Buglio, R., Andrews, M. (2014). *Big data beyond the hype*. [pdf]. Retrieved from https://www.ibm.com/developerworks/community/blogs/SusanVisser/entry/big_data_beyond_the_hype_a_guide_to_conversations_for_today_s_data_center?lang=en
- 451 Research. (2015). *Data platforms map*. [diagram]. Retrieved from <https://451research.com/dashboard/dpa>