

# Insights2Success

The way of business solutions

www.insightssuccess.com  
November 2017

THE  
10  
BEST

**HADOOP**

SOLUTION PROVIDER  
COMPANIES 2017

## EXPERT'S COLUMN

THE POWER OF  
PREDICTIVE INSIGHTS

## INDUSTRY FORECAST

HADOOP MARKET WILL  
UPSURGE TO \$54.2 BILLION  
IN THE NEAR FUTURE

## TECH-TALK

TRAJECTORY OF HADOOP'S  
FUTURE AND ITS EFFECTS  
ON THE INDUSTRY

DAVID FRIEDLAND  
COO

# IRI Voracity

An Insatiable Appetite for Data

# Editorial




## An underrated technology with the capability to change the world

In the time period 2012 to 2014, Hadoop started taking over the market of Big Data, when the industry was already marked by wave of mergers, acquisitions and financial rounds with high level valuations. The scenario looks entirely different today. The industry experts say that Hadoop is the only cost-sensible and scalable open source alternative to the other commercially available Big Data management packages. The technology is now an integral part of almost any commercially available Big Data solution and also the de-facto industry standard for business intelligence (BI).

Even though the usability and essentiality of this technology is incomparable, Hadoop has failed to deliver enough in terms of revues by the year 2015. The industry reports suggest that during the duration of three years from 2012 to 2015, venture capital, acquisition money and R&D budgets were playing the leading roles in the growth and development of the technology. The industry saw some extraordinary Hadoop trends like rapid shift from batch processing to online processing, emergence of MapReduce alternatives like Spark, Storm and DataTorrent.

Like any other technology, Hadoop has been through obvious setbacks but it seems that the technology is here to stay and grow with the rapid advancements in the near future. The foreseers of the industry say that there will be commercially supported Hadoop-related software, hardware and appliances helping enterprises in consulting, integration, middleware and support as well as in training and outsourcing.

Not only in the sphere of Information Technology and business world, Hadoop is one such technology which is redefining terms even in the healthcare industry. An independent research and survey has revealed that the technology has played an active role in reducing cost overhead, curing diseases, improving profits, predicting epidemics and enhancing the quality of human life by preventing deaths. This technology is now being adopted in the scientific research labs, hospitals and other medical institutions to reduce healthcare costs by changing the models of treatment.

Even the technology was not making sufficient profits in the industry for a long time, there apparently seems no reason to underrate or underestimate its capabilities. Ever since its invention, Hadoop has been redefining terms in the technological world and helping businesses improve. It clearly seems that the technology is nothing without the contribution of Hadoop solution providers. With time, people will surely understand the capabilities of the technology and its advancements will surely bring a change in the world for good. 

*Rajarshi Chatterjee*  
Rajarshi Chatterjee

COVER  
STORY



08

# IRI VORACITY

AN INSATIABLE APPETITE FOR DATA

## INDUSTRY FORECAST

Hadoop Market will upsurge to \$54.2 billion in the near future

32

ARTICLE

## TECH-TALK

Trajectory of Hadoop's Future and its Effects on the Industry

20

## EDITOR'S PICK

Big Data Analytics Key To Business Growth

38

## INDUSTRY INSIGHT

An Insight on the Opportunities in the Big Data Industry

26



## Serving Technology to Help Clients’ Improve Businesses

**T**he history of Hadoop goes back to the early 21<sup>st</sup> Century. Hadoop was invented by the famous advocate and creator Doug Cutting, he also invented Apache Lucene and Apache Nutch, In the mid 21<sup>st</sup> century when the technology started getting into the veins of the world of information technology, more and more companies started adopting the technology to improve their services. In the year 2008, Yahoo! announced that its production search index was being generated by a 10,000-core Hadoop cluster. In the same year, the technology was made its own top-level project at Apache which further confirmed its success and also its diverse and active community.

Interestingly, again, in the same year, Hadoop created a milestone in the technology sector by breaking all world records to become the fastest system to sort a terabyte of data.

Hadoop is one such technology which is now getting adopted by businesses all over the globe and is helping enterprises leveraging their growth. To showcase the companies which are enriching the industry with extraordinary technology, we have come up with the issue on “**The 10 Best Hadoop Solution Provider Companies 2017.**”

On the cover of this edition, we have, “*IRI Voracity: An Insatiable Appetite for Data*”. IRI is a leading big data software and solutions provider with both tools and a platform that address the data processing, protection, and provisioning needs of enterprises large and small. It is a data management ISV founded in 1978, serving as a centralized data marshaling area and one-stop solution stack for data discovery, integration, migration, governance, and analytics. IRI touts Voracity as the “*only affordable, high-speed platform for managing data in flat files, DBs, HDFS, and cloud apps, from profiling to presentation.*”

Along with this interesting Cover Story, we have enlisted some of the major contributors of the field, one of that is, “*PSSC Labs: Delivering Hand-Crafted HPC & Extraordinary Big Data Computing Solutions*”. To expand the potentials and show how businesses can benefit with the Hadoop technology, PSSC Labs comes into the scenario with over 25 years of experience in delivering solutions ensuring relentless performance with the absolute lowest total cost of ownership. The company has set an exceptional record by offering companies custom-built options outside of expensive proprietary designs offered by larger solution providers—an option once only considered variable—for enterprises with massive infrastructure needs.

To add an interesting angle to the edition, we have come up with some knowledge enriching articles from the experts of the industry like “*The Power of Predictive Insights*” by **Steven Ramirez, CEO at Beyond the Arc, Inc.** Along with such amazing pieces to read about Hadoop, we have also covered some of the amazing articles crafted by our in-house editors as *Trajectory of Hadoop’s Future and its Effects on the Industry* and *Hadoop Market will upsurge to \$54.2 billion in the near future: Industry Forecast*, among many others.

Hope this issue will leave a mark on your mind and will enrich your knowledge about Hadoop technology. 

Cover Story

THE 10 BEST   
**HADOOP**  
SOLUTION PROVIDER COMPANIES 2017




David Friedland  
COO

# IRI




# VORACITY

An Insatiable Appetite for Data



Voracity is designed to support the full scope of data-driven work in a familiar place ... a single pane of glass where users can find, process, govern, and analyze their data, plus manage their jobs and metadata



“

VORACITY'S SEAMLESS INTERCHANGEABILITY OF ENGINES MEANS THAT DATA ANALYSTS, ETL ARCHITECTS AND GOVERNANCE TEAMS CAN LEVERAGE THE SAME ECLIPSE PANE-OF-GLASS TO DESIGN AND RUN JOBS. THERE'S NO NEED TO LEARN HADOOP CODE

”

In the era of big data, two of the biggest challenges that IT professionals face are: 1) *speeding* insight from variable data sizes, formats, and streams, and 2) *securing* personally identifying information (PII) to protect corporate reputations and to comply with data privacy laws. Some try to address the performance challenge by powering legacy data integration or virtualization suites with huge servers. Others try complex Hadoop programs or unfamiliar database technologies. For data security, they turn to costly classification and de-identification technologies, and specialized compliance experts.

Many continue to seek out proven solutions they can afford. **IRI (Innovative Routines International, Inc.)**, is a data management ISV founded in 1978 focused on fast, feature-rich processing and protection technology for data big and small.

According to IRI's SVP and COO David Friedland, the company's early roots were in moving mainframe sort/merge/report jobs into CP/M, DOS, UNIX, and Windows. This initial mission led IRI to develop and parallelize more data mapping functions, which also made its "CoSort" product popular with DW and BI architects needing a faster ETL and data preparation engine.

As IRI grew in big data processing markets (long before Hadoop emerged), it developed even more data-centric capabilities for profiling, processing, protection, presentation, and prototyping, including: data searching and classification, data integration and replication, data masking and encryption, data cleansing and reporting, and test data generation. Today, IRI delivers eight (8) data management and protection software products, which are supported in more than 40 international offices.

**Big Data Manipulation in a Managed Environment**  
CoSort is the default data processing engine in IRI's modern "total data management" platform, Voracity. The platform can also use Hadoop engines too, but more on that later.

The purpose of Voracity is to be a centralized data marshalling area and one-stop solution stack for data discovery, integration, migration, governance, and analytics. IRI touts Voracity as the "*only affordable, high-speed platform for managing data in flat files, DBs, HDFS, and cloud apps, from profiling to presentation.*"

Voracity uses a popular (and free) graphical integrated development environment (IDE) called IRI Workbench. Because it is built on Eclipse, the GUI for Voracity is automatically familiar to millions of users, and is a fully extensible solution stack. Many free and commercial plug-ins can open in Voracity's user workspaces and run within Voracity workflows.

Using these flexible Eclipse "workspaces," different stakeholders can work alone or in teams to profile and classify, integrate and harmonize, clean and mask, prototype or replicate, and blend or analyze their data, as well as track its changes through time.

More specifically, Voracity performs multiple functions within five key data management areas:

- *Data Discovery* -- search, extract, structure, profile, classify, and diagram data sets
- *Data Integration* -- extract, transform, load (ETL), change data capture, pivoting, etc.
- *Data Migration* -- data type, file format, endian, and database conversion or replication
- *Data Governance* -- cleansing, masking, test data,



static and streaming data -- from files, DBs, IoT, and more -- without Apache-project complexity or mega-vendor costs.

**Development Background**

IRI had been building toward Voracity long before realizing it. But once dedicated to the platform, IRI began to combine its existing products with other state-of-the-art technologies.

Years of internal and external innovation later, Voracity is now more than the sum of its parts. For example, the platform added searching, profiling, and classification wizards to discover data ahead of analytic, quality, and masking operations.

Voracity also supports newer data delivery methods like Kafka and formats like JSON along with old school COBOL files and relational databases. It combines IRI's proprietary engines with open source communication and parsing protocols to address all the challenges of big data today: volume, variety, velocity, veracity, and value.

- *Analytics* -- embedded reporting, BIRT and dashboard integration, or data wrangling

This diagram provides a view of how data flows through typical work sequences:

**Fast Data Munging and Masking (With or Without Hadoop)**

Many data stores are now in Hadoop Distributed File Systems (HDFS), and the cost of cluster hardware continues to fall. As a result, there's an increasing need to process and protect data in HDFS files, Hive, etc. With this comes the challenge of a steep learning curve for Hadoop users.

IRI addresses these issues in the platform. The data transformation, masking, and reformatting jobs built visually for CoSort can also run automatically in Hadoop MapReduce 2, Spark, Spark Stream, Storm, or Tez.

*"Voracity's seamless interchangeability of engines means that data analysts, ETL architects and*

**With Voracity, you don't need a PhD to use it, a fat purse to buy it, or large team to support it**

*governance teams can leverage the same Eclipse pane-of-glass to design and run jobs. There's no need to learn Hadoop code."*

IRI sees the convergence of Hadoop commoditization and Voracity task consolidation as the perfect opportunity for smaller and mid-sized companies to capitalize on big data. IRI says that Voracity can manage both

A key design goal was also accessibility. IRI built Voracity to be user-friendly for a wide range of groups, including BI/DW architects, data scientists, DBAs, and GRC

Officers. Another goal was to future-proof it against change. The versatility of Voracity's programs and constant evolution in Eclipse functionality make the platform a chameleon for multiple tasks today, and future data processing requirements tomorrow.

Many thought leaders in the data management industry contributed to Voracity, guiding it during its development and positioning stages to make sure it hit its mark. Analysts at Gartner, consultants at Athena Solutions, Big Data Dimension, and the Data Governance Institute all weighed in on the platform, as did the inventors of AnalytiX DS Mapping Manager and the Data Vault (Dan Linstedt). All of them ensured that Voracity became a flexible, outcome-driven platform.

#### Customer Adoption

Although IRI continues to expand, it already has many multinational customers. Banks and insurance companies like *Bank of America*, *AIG*, *HSBC*, and *AXA* process their data with Voracity's CoSort engine, as do airlines like *American*, *Japan*, and *Lufthansa*, and automotive companies like *Hyundai*, *Nissan*, and *Mercedes-Benz*. Hosts of other conglomerates like *Visa*, *Nestle*, *Samsung*, *Capgemini*, *Accenture*, *Rolex*, *Sony*, and *The Walt Disney Company* also rely on IRI's data manipulation technology.

*"We are proud to have thousands of users worldwide using IRI software in contexts like data integration and data masking, which leading publications and analysts firms like Gartner, IDC, and The Bloor Group all recognize,"* David remarked.

#### Building Partnerships to Build Business

IRI routinely collaborates with specialty providers who contribute to the Voracity ecosystem. By blending into the IRI data fabric in Eclipse, these developers can add their value to the platform with minimal user impact.

To illustrate, IRI recently announced a partnership with AnalytiX DS to enhance Voracity's metadata management capabilities for ETL and data quality users. The companies unveiled their complementary functions in adjacent exhibits at the Dataversity Enterprise Data World conference and expo in Atlanta this year.

While talking about the partnership, David said, *"Both companies are excited about bringing the combined technologies to market under a single, integrated offering."* He commented on the benefits that their clients will be getting and said, *"The bridge between our platforms is built on an API-level integration of metadata. This enables anyone with either stack to use the strengths of the other on a pay-to-play basis."*

#### Staying Ahead of the Data Management Industry

IRI attributes its technical success to: 1) an organically-grown code base focused on big data processing speed; 2) simple and open metadata; 3) the extensibility of Voracity's Eclipse design/deployment GUI; and, 4) input from industry thought leaders.

IRI also credits its success to conservative growth, relatively-low marketing overhead, and the priority given to its most loyal users in feature-function decisions, support resources, and licensing flexibility.



Voracity is a total data management solution stack that bends ETL mega-vendor cost curves, eliminates multi-tool complexity, and mitigates Apache project risks



This combination of factors allow IRI to deliver enterprise-class performance and functionality at affordable subscription prices.

Moving into its next 40 years, IRI sees Voracity as a key to the growth of the company and the industry. IRI continually ranks among the data management industry's top firms. Database Trends and Applications (DBTA) ranked the tools in IRI's Data Protector Suite -- also key components of Voracity -- as a *Trend Setting Product* in 2017. *CV Magazine* named Voracity the *Most Price-Performant Big Data Management Platform*.

#### Still Active Founders

**Paul Friedland, CEO**, started the company in 1978. Paul's innovation in high-performance, high-volume data processing began even earlier when he was cited in Knuth's *Sorting and Searching*. Through decades of continuing innovation in co-routine architecture, multi-threading, data manipulation, and task consolidation, he made IRI a leader in the big data processing industry long before Hadoop was introduced. His son, **David Friedland, COO**, joined IRI in 1998 after working in technology marketing and international journalism. Today, he manages both partner and product line growth, works with analysts and stakeholders, blogs on technical topics, and speaks at trade conferences.

David said, *"Despite the rich technical history I've seen at IRI, the years ahead bode even better. We are in the middle of exponential data growth. Voracity's myriad solutions in data governance and analytic-related applications coupled with its price-performance position in Hadoop-fueled markets bring us new opportunities every day."*

While discussing the future of the company, he added *"I see us leveraging partner technologies for NLP, machine learning, and rich visualizations as a way to add even more value to the Voracity platform going forward."*

Voracity is designed to support the full scope of data-related work in a familiar place to govern data, metadata, and users

