TDWI RESEARCH

TDWI BEST PRACTICES REPORT

Hadoop for the Enterprise

Making data management massively scalable, agile, feature-rich, and cost-effective

Philip Russom

TDWI Research Director for Data Management

April 14, 2015

TDWI would like to thank the following companies
for sponsoring the 2015 TDWI Best Practices research report:

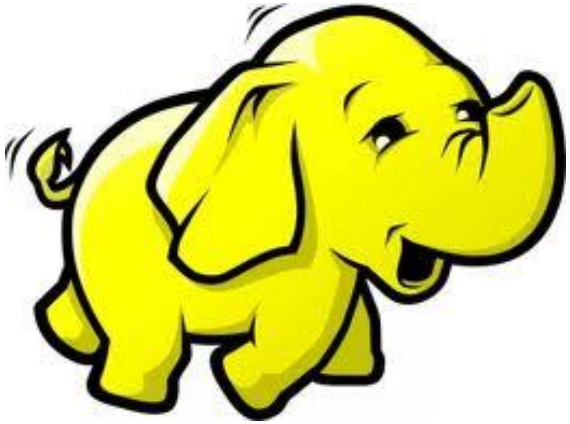
Hadoop for the Enterprise

This presentation is based on the findings of that report.

STAY TUNED – At the end of this webinar, learn how to download a free copy of the report.

Agenda

- Hadoop for the Enterprise
  - *New Directions for Hadoop*
  - *Drivers of Change*
- Why bother?
  - *Opportunities & Innovations*
  - *Benefits & Barriers*
- Misc. Hadoop Metrics
  - *Hadoop in Data Architectures*
  - *Trends in Hadoop Use*
- Top Ten Priorities for Enterprise Hadoop

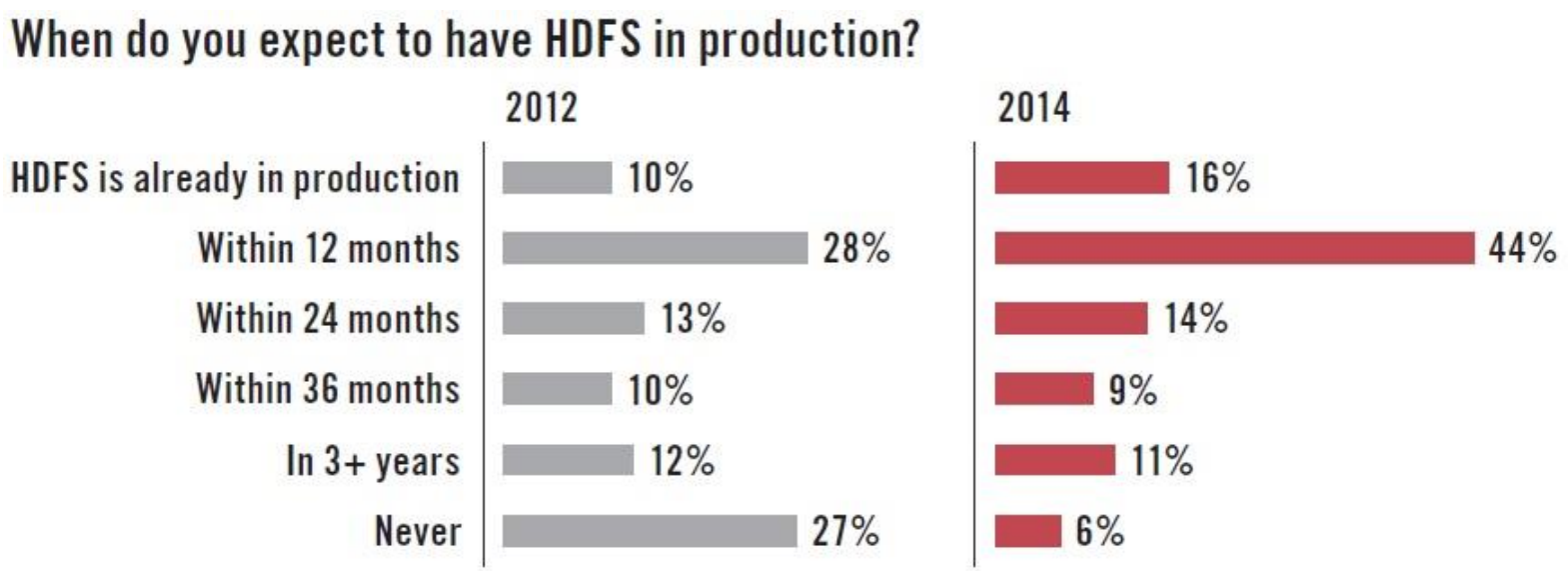


PLEASE TWEET
@pRussom, #TDWI, #EDW, #DataWarehouse, #DataArchitecture, #Analytics, #Hadoop

We assume you're familiar with Hadoop. If not, read the TDWI report "Integrating Hadoop with BI and DW."

User Adoption of Hadoop is Up

- Hadoop clusters in production are up
  - *Up 6 percentage points, which is a 60% gain over two years*
- Many more Hadoop clusters will come online soon
  - *60% of respondents will be in production by Q1 2016, which is a huge leap forward from the current 16%*
- Very few users surveyed have ruled out Hadoop
  - *Rising commitment: 27% in 2012, down to 6% in 2014*



Source: TDWI surveys.
Based on 263 respondents in 2012; 247 in 2014.

Enterprise Hadoop is Evolving as it Grows

- Hadoop is expanding
  - *More industries*
  - *Use cases*
  - *Enterprise breadth*
- Hadoop changes as it goes enterprise scope
  - *Staffing*
  - *Development methods*
  - *Economics*
  - *Ownership*
- Hadoop clusters are becoming enterprise assets
  - *Shared IT infrastructure*
  - *Still departmental in some cases*

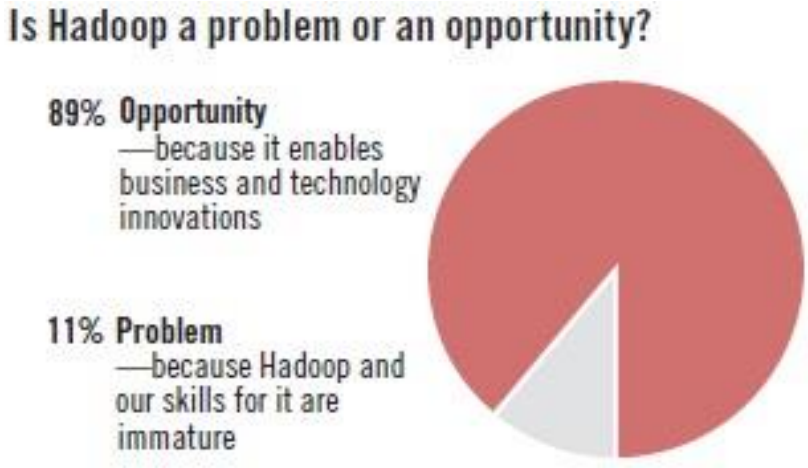
Hadoop's Use Cases are Evolving

- Internet firm operations
  - *Extreme big data handling, reporting, and analytics*
- Extensions of the BI/DW technology stack
  - *Data warehouse scalability, exotic data handling*
  - *Analytics with algorithms, sets, exotic data*
  - *Data integration cost reduction, scalable staging*
- Data archiving
  - *Online, active archive*
  - *Scalable and cost effective*
- Content management
  - *Email archives*
  - *Records management*
  - *Document management*

Drivers for User Adoption

- Hadoop for the Enterprise **technical** drivers:
  - *Scalability*
  - *Low cost*
  - *Many data types*
- Hadoop for the Enterprise **business** drivers:
  - *Analytics*
  - *Data exploration*
  - *value from Big Data*

Hadoop Enables Business and Technology Innovations



- Collocate multi-terabtye datasets on Hadoop
  - *Enables unprecedented data exploration and discovery, which in turn reveals top- and bottom-line opportunities*
- Some Hadoop-driven innovations are incremental
  - *Expand the data samples of existing customer analytics, which elevates customer service, retention, account growth*
- Hadoop excels with streaming and exotic data types
  - *Provides visibility and insights into business processes and entities that were previously dark*

Source: TDWI survey run in late 2014.
Based on 247 respondents.

Benefits of Enterprise Hadoop

In priority order, based on survey responses

- Advanced analytics
- Data warehousing and data integration
- Data scalability
- New and exotic data types
- Business applications

Source: TDWI survey run in late 2014. Based on 1348 responses from 247 respondents.

Barriers to Enterprise Hadoop

In priority order, based on survey responses

- Lack of skills
- Weak business support
- Limited security
- Immature tools
- High personnel costs for data scientists, programmers, consultants

Source: TDWI survey run in late 2014. Based on 1162 responses from 247 respondents.

MISC HADOOP METRICS

Titles, Training, and Tools

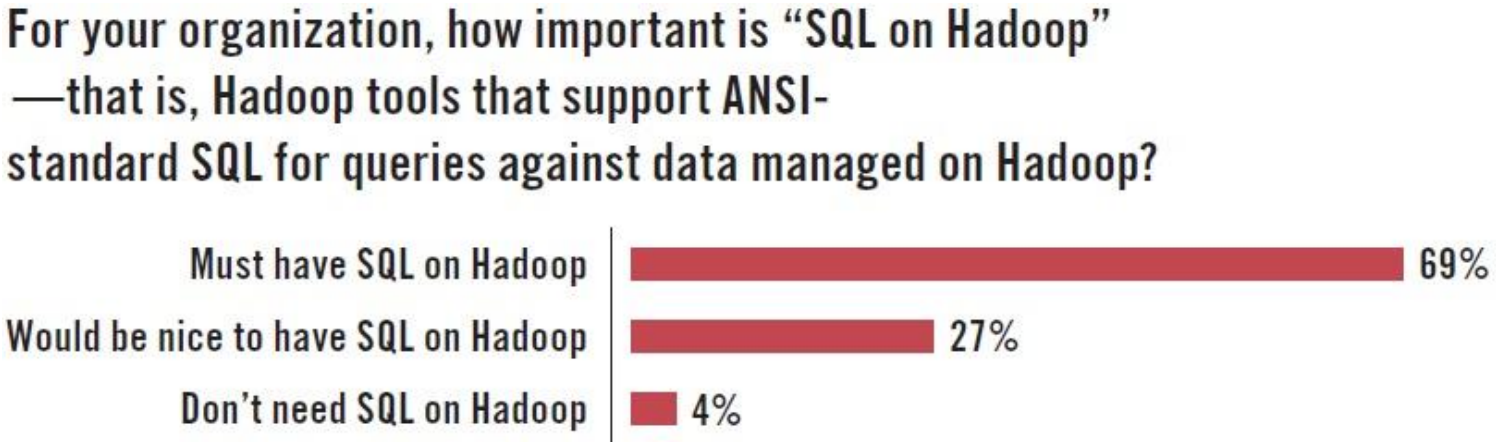
- Hadoop job titles, in survey priority order
  - *Data Scientist tops the list at 16%*
  - *Architect, analyst, or developer (37%)*
    - Whether specializing in data or apps
  - *New title: Big Data Specialist at 8%*
- Firms train employees in Hadoop (73%)
  - *They can't find or afford experienced folks.*
- 58% of Hadoop development is done with a mix of hand-coded programs & high-level tools
  - *23% hand-coded only*
  - *14% tools only*
  - *5% other*

Making Hadoop “Enterprise Grade”

- Purely open-source Hadoop is missing items usually considered enterprise requirements
  - *Multiple approaches to security*
    - Apache Hadoop is limited to file-permission checks and access control based on Kerberos
    - Mature IT depts want multiple approaches: roles, directories, encryption, masking…
  - *Better administrative and development tools*
    - Still immature compared to rest of market
  - *Maintenance and support*
    - IT needs a “single throat to choke”
- Vendor distributions of Hadoop – or “distros”
  - *Enterprise users are trending toward distros*
  - *Distros usually include additional tools for security and admin, plus maintenance/support and sometimes consulting*

SQL and other Relational Functions

- Data professionals want and depend on SQL
  - *It must be ANSI standard, low latency, low cost*
- SQL on Hadoop versus SQL off Hadoop argument
  - *Users interviewed want BOTH !*
  - *In survey, SQL on Hadoop is a "must have" (69%)*
  - *Only 4% don't need SQL on Hadoop*



Source: TDWI survey run in late 2014.
Based 99 respondents.

Evolving Data Mgt Best Practices and Hadoop

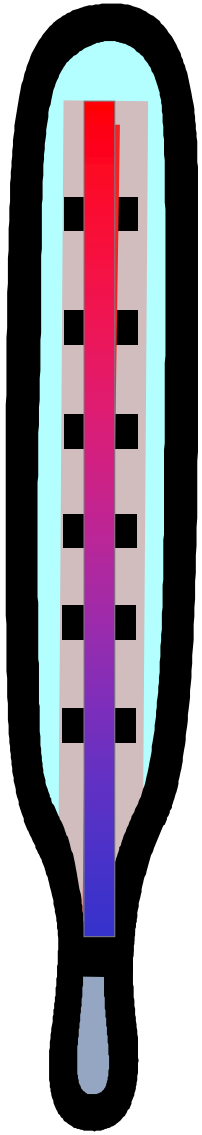
- Some practices for reporting, warehousing, data integration:
  - *Early processing – ETL, staging, merging, transformations, modeling, cleansing, standardization, depupe, sort…*
  - *Late ingestion – after much processing, data is loaded into a DW or similar database. Its state is squeaky clean, but for a single purpose.*
- Some practices for discovery analytics or time-sensitive data:
  - *Early ingestion – source data, in its raw extracted state, is loaded immediately into Hadoop or some other target*
  - *Late processing – on an "as needed" basis, source data is processed for reporting and analytics. Can be repurposed many ways, any time.*
- You can do either or both on Hadoop, depending on app requirements
  - *Even so, early ingestion/late processing is more common*
    - Hadoop is often used as a computational platform for advanced analytics, which evolves constantly, so late processing is preferred
    - Early ingestion is a practical strategy for coping with large data volumes or streaming data, as is typical of some big data types

Hadoop integrated with a Relational DBMS

The strengths of one balance the weaknesses of the other

- A Relational DBMS is good at:
  - *Metadata management*
  - *Complex query optimization*
  - *Query federation*
  - *Table joins, views, keys, etc.*
  - *Security, including roles, directories*
  - *Much more mature development tools*

- HDFS & other Hadoop tools are good at:
  - *Massive, linear scalability*
  - *Lower cost than most DW platforms & analytic DBMSs*
  - *Multi-structured & no-schema data, not just structured & relational*
  - *ETL and ELT functions that are not inherently relational*
  - *Custom code for algorithmic analytics*



Hot Stuff, Today & Tomorrow

Based on responses to the report survey

- Top tool types integrated with Hadoop today
  - *BI and DW uses – reporting (36%), DW (36%), analytics (32%), visualization (30%), data integration (29%)…*
- Top use cases by projected growth
  - *Ent. data hubs (9% growth), archiving (7%), BI (5%), DW (4%), adv'd analytics (4%), operational apps (4%), content mgt (4%)…*
- Top growth: data quality (DQ) & master data mgt (MDM)
  - *Integrated with Hadoop Today – DQ=11%; MDM=10%*
  - *Integrated with Hadoop in Three years – DQ=55%; MDM=45%*
- Top Hadoop tools from open source and similar origins
  - *Today – the "big five": HDFS, MapReduce, Pig, HBase, Hive*
  - *Today – the utilities: Zookeeper and Hue*
  - *Future growth – Real time: Spark, Storm; Near time: Impala, Drill*

Top Ten Priorities for Enterprise Hadoop

These are recommendations, requirements, or rules that can guide you.

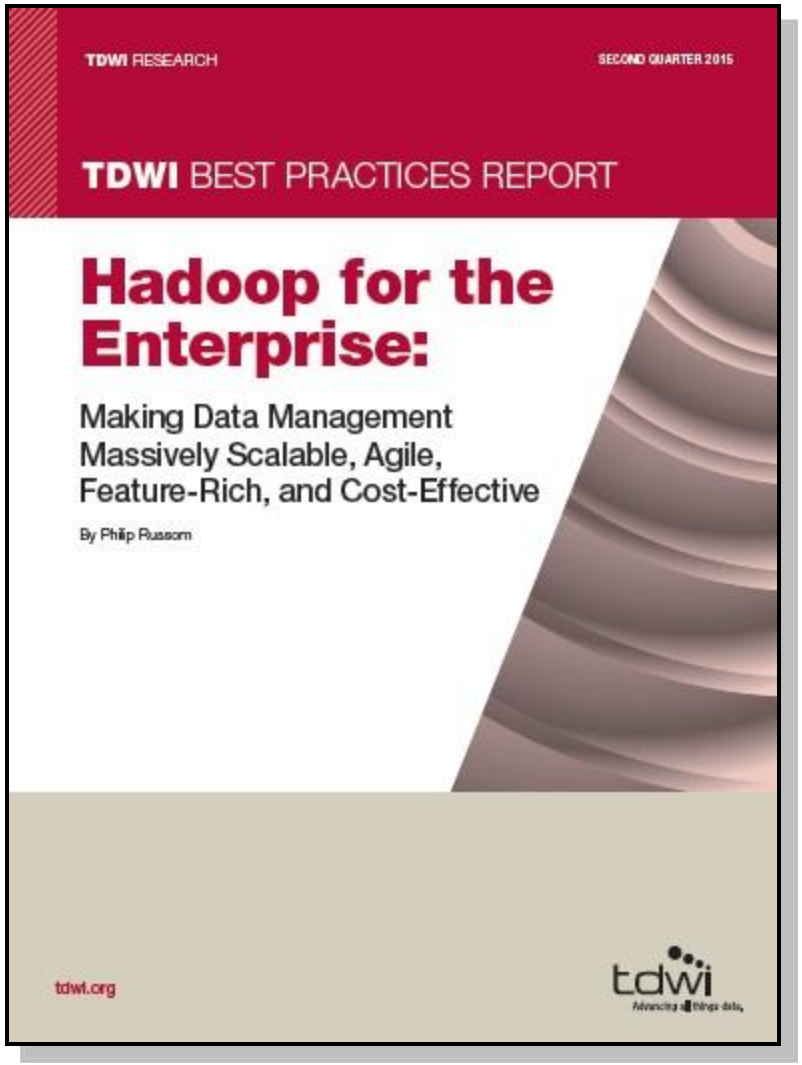
1. Be open to Hadoop and other new options.
2. Innovate with big data on enterprise Hadoop.
3. Base Hadoop adoption on business and technology requirements.
4. Know the hurdles, so you can leap over them.
5. Get training (and maybe new staff) for Hadoop and big data mgt.
6. Co-opt Hadoop to rethink the economics of data and content architectures.
7. Prepare for hybrid data ecosystems by defining places for Hadoop in their architectures.
8. Consider Hadoop use cases outside the usual BI/DW and analytic applications.
9. Make Hadoop data look relational, when needed.
10. Develop and apply a strategy for enterprise Hadoop.

Download a free copy of the report that this Webinar is based on

- Download the report in a PDF file at:

  tdwi.org/bpreports

- Feel free to distribute the PDF file of any TDWI Best Practices Report



Q&A

Philip Russom

Research Director for Data Mgt

TDWI

pRussom@TDWI.org

@pRussom on Twitter

linkedin.com/in/philiprussom