TDWI INFOGRAPHIC

Q1 2017 TDWI BEST PRACTICES REPORT
DATA LAKES: PURPOSES, PRACTICES, PATTERNS, AND PLATFORMS

When designed well, a data lake is an effective data-driven design pattern for capturing a wide range of data types, both old and new, at large scale. Organizations are adopting the data lake design pattern (whether on Hadoop or a relational database) because lakes provision the kind of raw data that users need for data exploration and discovery-oriented forms of advanced analytics. In the recent *TDWI Best Practices Report: Data Lakes – Purposes, Practices, Patterns, and Platforms*, we define data lake types, discuss emerging best practices, and take a look at user trends and readiness for data lakes. Here are several of the key survey results.

BENEFITS

DATA LAKES HAVE THE POTENTIAL TO MODERNIZE EXISTING DATA ECOSYSTEMS AND EXPAND ANALYTICS PROGRAMS.

If your organization were to implement Hadoop-based data lakes, which of the following use cases would most likely benefit?

49% ADVANCED ANALYTICS

49% DATA EXPLORATION AND DISCOVERY

45% BIG DATA SOURCE FOR ANALYTICS

39% EXTENSION OF DATA WAREHOUSE

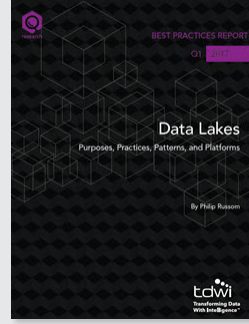**36%** DATA LANDING AND STAGING FOR DATA WAREHOUSING

BARRIERS

DATA LAKES ARE HARD TO SECURE AND GOVERN, AND HADOOP SKILLS ARE IMMATURE.

In your organization, what are the most likely barriers to implementing Hadoop-based data lakes?

41% LACK OF DATA GOVERNANCE

32% LACK OF DATA INTEGRATION TOOLS AND SKILLS FOR HADOOP

32% INADEQUATE SKILLS FOR BIG DATA

32% INADEQUATE SKILLS FOR HADOOP

31% LACK OF COMPELLING BUSINESS CASE

Source: Figures 4 and 5, pages 12–14 (Respondents could make up to six selections)

DATA LAKE ADOPTION

NEARLY A QUARTER OF RESPONDENTS ALREADY HAVE A DATA LAKE IN PRODUCTION, CONFIRMATION THAT LAKES WORK ON A TECHNICAL AND BUSINESS LEVEL.

When do you expect to have a data lake in production?

- **7%** Never
- **21%** In 3+ years
- **10%** Within 36 months
- **15%** Within 24 months
- **24%** Within 12 months
- **23%** Already in production
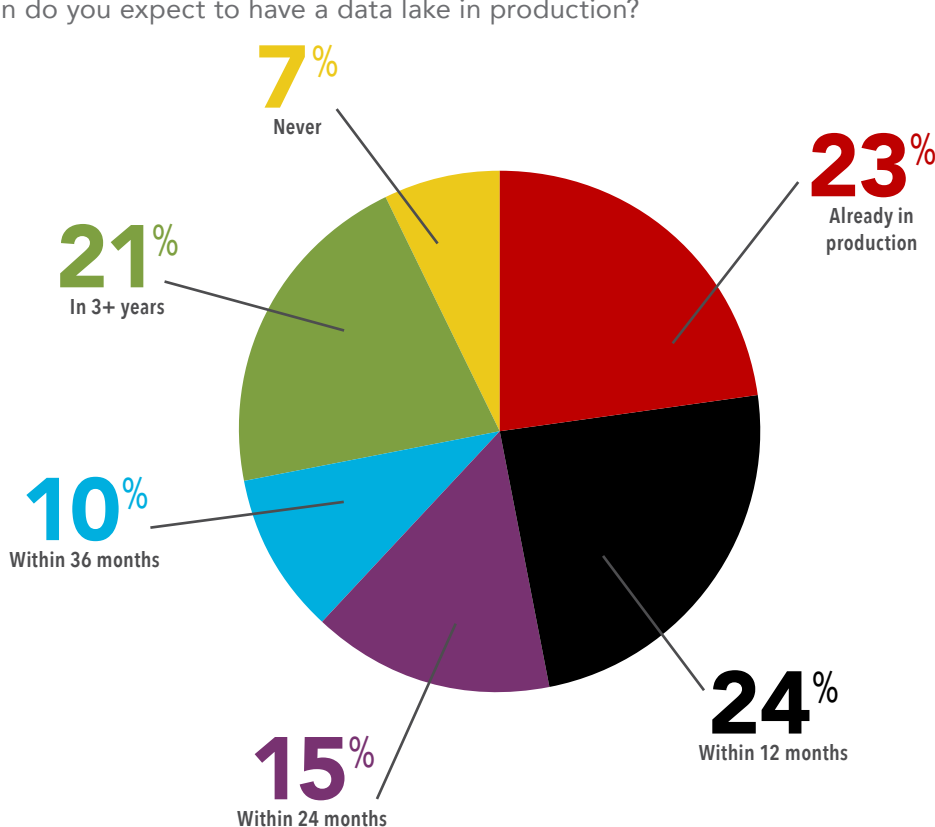
Source: Figure 8, page 16

PLATFORM ISSUES: HADOOP VS. RELATIONAL DATABASES

FOR RESPONDENTS WITH DATA LAKE EXPERIENCE, HADOOP IS PREFERRED OVER RELATIONAL DATABASES FOR MANAGING DATA'S EXPLODING DIVERSITY AND SCALE.
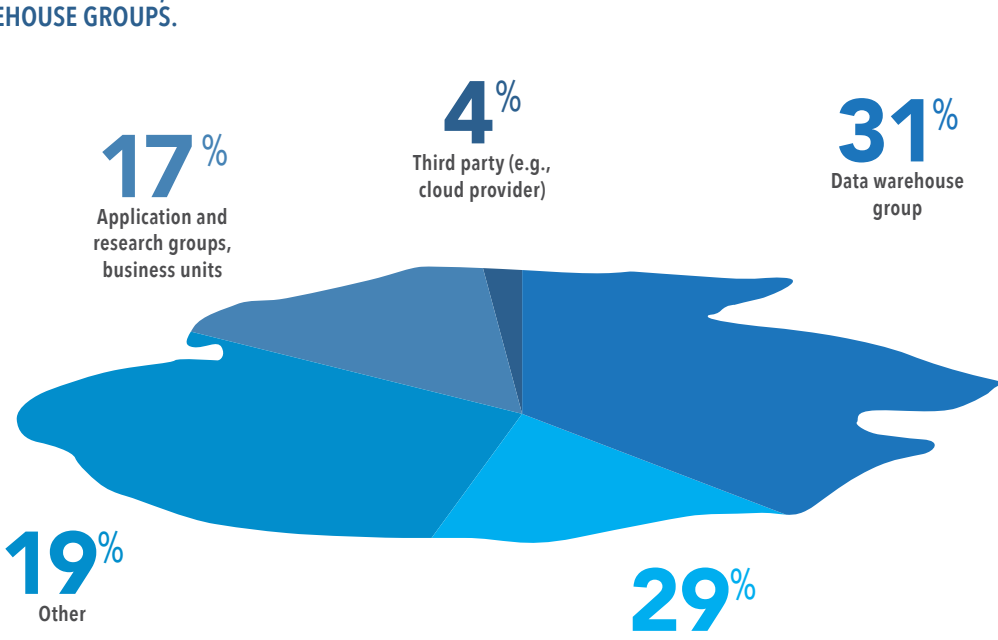
For the data lake you use most, what type of data platform is it deployed on?

- **53%** Hadoop
- **25%** Both
- **16%** Other
- **6%** Relational database management system
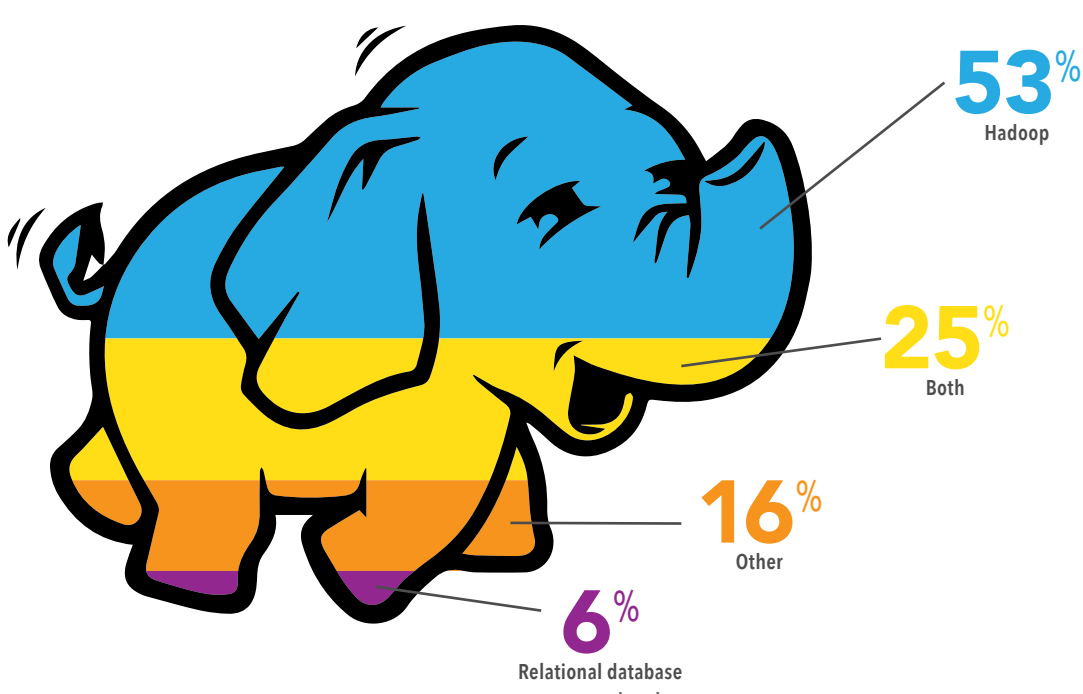
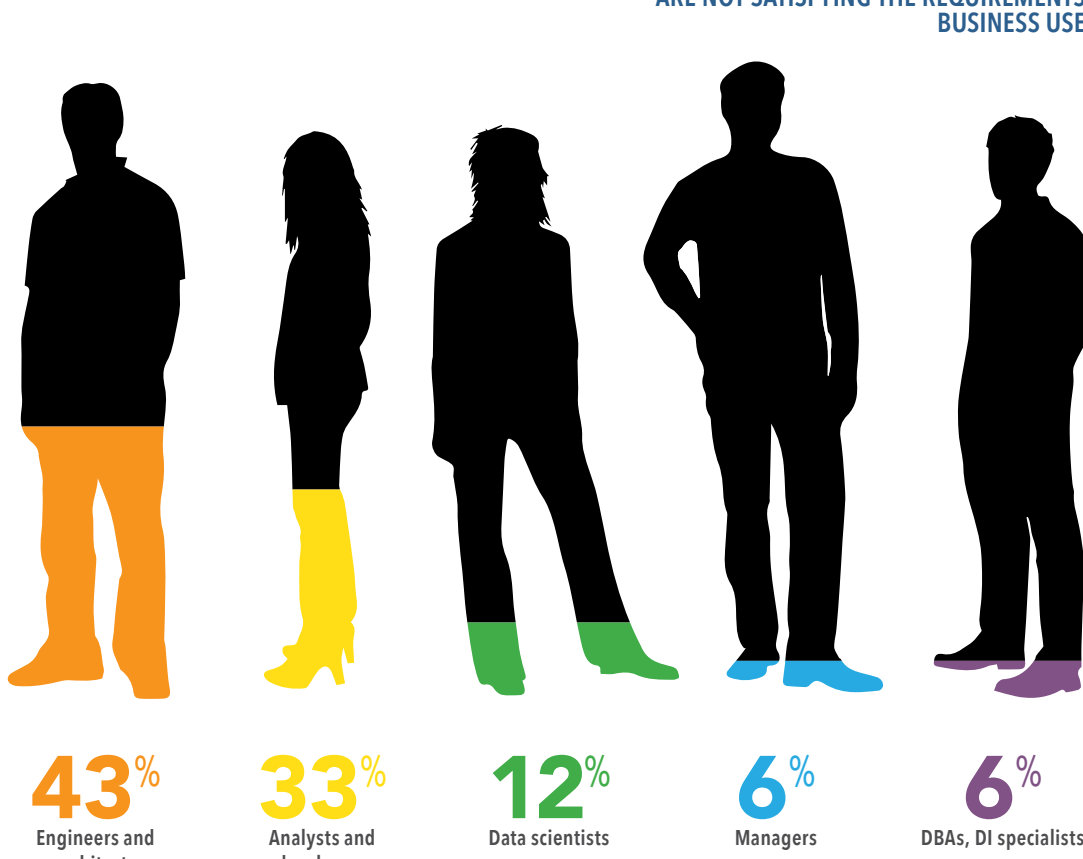Source: Figure 17, page 31 (Based on 75 respondents who have data lake experience)

DATA LAKE OWNERS

BECAUSE MOST DATA LAKES ARE DEPLOYED ON HADOOP, THE OWNERSHIP OF DATA LAKES IS SIMILAR TO HADOOP, WITH ADOPTION LED BY DATA WAREHOUSE GROUPS.

- **17%** Application and research groups, business units
- **4%** Third party (e.g., cloud provider)
- **31%** Data warehouse group
- **19%** Other
- **29%** Central IT

DATA LAKE WORKERS

DATA LAKES ARE BUILT AND USED BY VERY TECHNICAL PEOPLE AND, ON AVERAGE, ARE NOT SATISFYING THE REQUIREMENTS OF BUSINESS USERS.
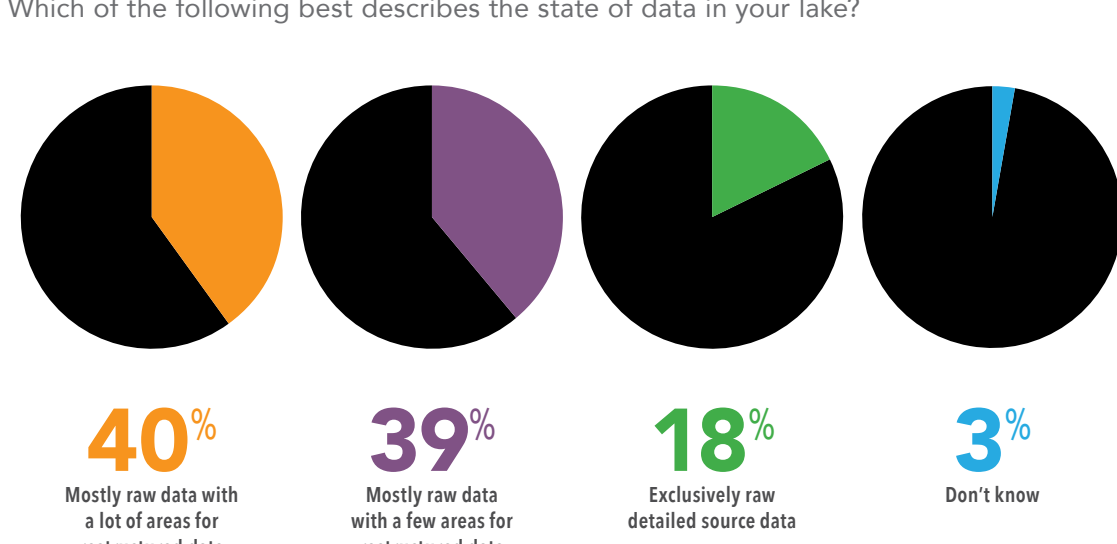

- **43%** Engineers and architects
- **33%** Analysts and developers
- **12%** Data scientists
- **6%** Managers
- **6%** DBAs, DI specialists

Source: Figures 14 and 15, pages 21–22 (Based on respondents who have data lake experience, see figures for details)

DATA CHARACTERISTICS IN DATA LAKES

ALTHOUGH DATA LAKES MANAGE MOSTLY RAW DATA, THE MAJORITY PRACTICE (79%) IS TO HAVE BOTH RAW SOURCE DATA AND AREAS DEVOTED TO STRUCTURED DATA.

Which of the following best describes the state of data in your lake?

- **40%** Mostly raw data with a lot of areas for restructured data
- **39%** Mostly raw data with a few areas for restructured data
- **18%** Exclusively raw detailed source data
- **3%** Don't know

Source: Figure 12, page 19 (Based on 72 respondents who have data lake experience)

BUSINESS USE CASES

DATA LAKES CONTRIBUTE TO TECHNOLOGY SCENARIOS BUT ALSO HAVE COMPELLING REAL-WORLD USE CASES.

- Enable exploration, discovery, and self-service
- Enable new analytics and expand old ones
- Can be extended to handle data in real time
- Enable fraud discovery and prevention

Source: "Compelling, Real-World Use Cases for Data Lakes," pages 8–9.