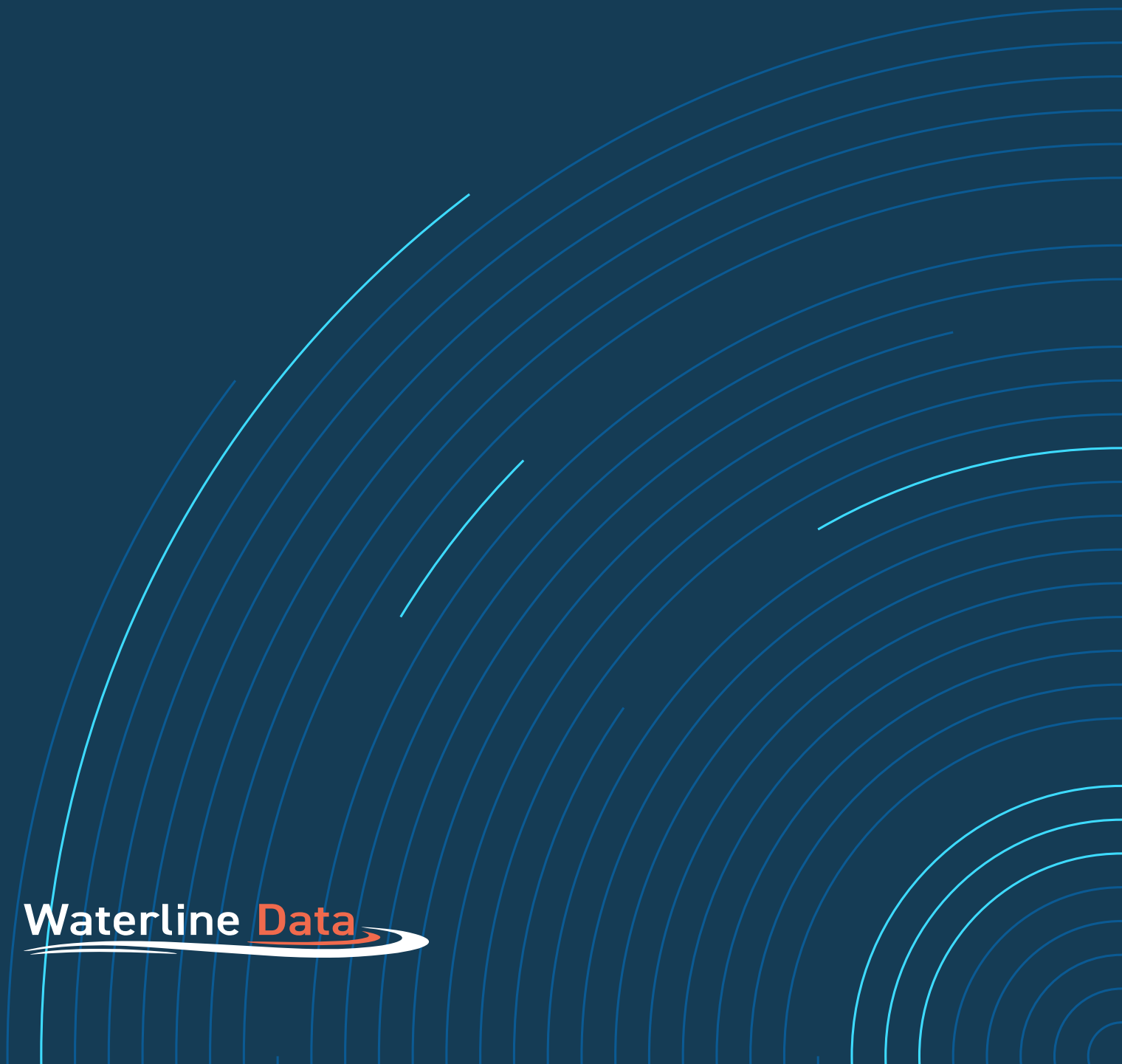
SOLUTION OVERVIEW

DATA CATALOGS FOR DATA RATIONALIZATION

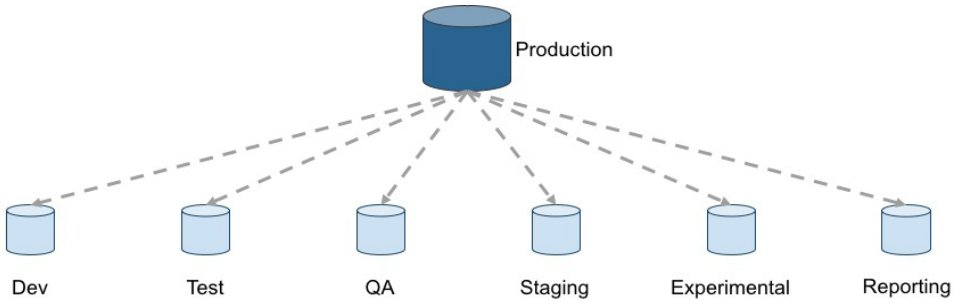


Waterline Data

Introduction

How big of a problem is data redundancy? If you are like most companies, it is much bigger than you would care to admit. For most businesses data has become like flies to flypaper.

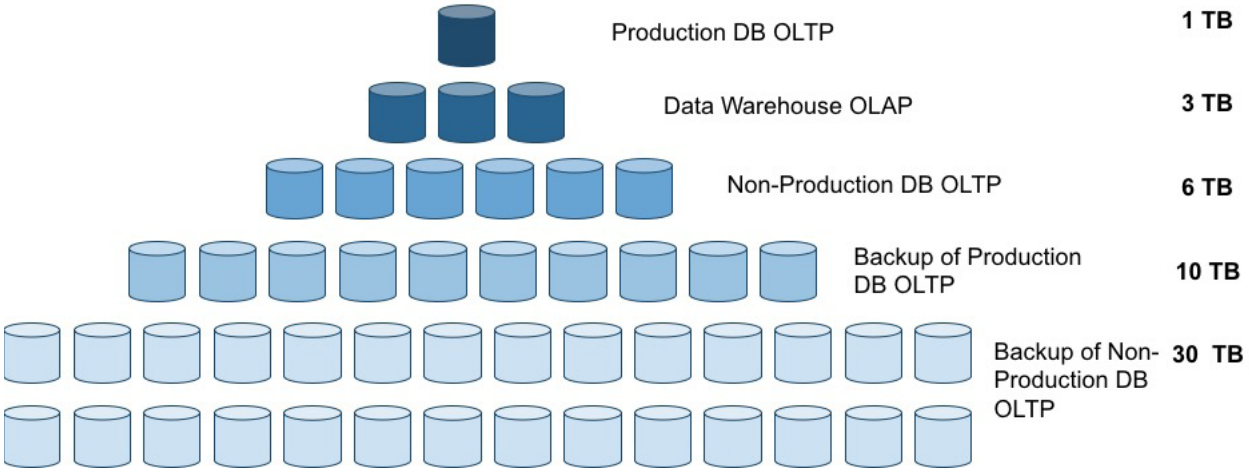
When you consider that for every production database, you also tend to spawn a development, test, QA, staging, experimental and reporting database, it doesn't take long for that single database to result in 6x more in total volume.



That additional volume requires additional database licenses, plus storage, as well as dat base administrators to manage the entire thing

.

In fact, if we think about the data proliferation pyramid, the multiplication factor is almost 50x . And once cost of storage began its steep descent in 2005, and cloud providers came along and began to provide lots of space at low cost, organizations used that as an excuse to continue to hoard data, even if they weren't actually using it.



The Hidden Cost of Big Data

In 2016, Veritas surveyed 10,000 organizations and 83% of IT decision makers said their organizations were “data hoarders.” This is neither a surprise nor a bad thing because the cost of storage has become relatively cheap. But while it is cheap, it still isn’t free.

83% of organizations are ‘data hoarders”

Let’s do the math!

In one example, a consumer retail company has over 8 petabytes of data currently costing them about $1 per gigabyte (this is a very conservative number that does not consider lots of factors including database software costs, people costs or storage management software), or $8 million to store. They also estimate that approximately 30% of that data (2.4 petabytes) is redundant and can be eliminated, resulting in savings of $2.4 million.

So, what is stopping them? They know at a high level they have lots of duplicate data, but at a granular level they don’t know exactly what information is the original source, what is replicated and what can be eliminated. They have no data map tracking all their data, its freshness or its provenance. The result is that they hoard everything, keeping redundant legacy development and QA data around for old projects that have long since been cancelled. They maintain data marts that were created for a specific one-time analysis, but then never delete them, all the time continuing to pay the $1 per GB for duplicate data that is no longer in use.

And these storage costs don’t even cover all the ancillary costs of redundant dark data. Much of that dark data could also pose security, compliance and other corporate risks that are hard to calculate. Imagine if some of that data violated the upcoming EU Governing Data Protection Regulation (GDPR), also known as the right for a consumer to be forgotten. That regulation carries a fine up to 4% of worldwide revenues.

The implication is that redundant, dark-data can no longer be left sitting around. The physical cost and the potential risk is becoming too high.

Using a Data Catalog to Rationalize Data

The resolution to this issue is conceptually simple. By implementing Waterline Smart Data Catalog, your organization can document the location, quality and lineage of data located in relational sources, in the cloud or in Hadoop clusters. A smart data catalog automatically tags the data fields across different data sets with common business terms, documents its overall quality, existing data lineage and inferred missing lineage. During this process, it also identifies where you have high levels of redundant data that can be rationalized.

the catalog identifies high levels of duplicate data for rationalization

How Does It Work?

The Waterline Data Catalog is built on the premise that there are three parts to the data cataloging process. The first is to help data professionals discover, organize and curate the data; the second is to allow governance professionals leverage tagged data for compliance reporting and access control and the third is to expose the newly organized data for business professionals to use.

Discover: Automatically and incrementally "fingerprint" data and infer data lineage at scale by analyzing actual data values

Organize: Uses machine learning to automatically tag and match data fingerprints to glossary terms

Curate: Human reviewers accept or reject tags, machine learning fine tunes the tagging process and improves the matching algorithm

Compliance: Map your compliance policies to your data assets: acceptable use, legal holds, expiry.

Reporting: Simplify ongoing mandated reporting as the catalog uncovers "dark data" and catalogs it dynamically

Access control: Automate data access control via tag based security

Search: Search for data through the Waterline GUI or through integration via 3rd party applications

Rate & Collaborate: Users collaborate to create subjective crowdsourced ratings/reviews which combined with objective profiling metadata provides users with a view into data quality and usefulness.

Benefits for Data Rationalization

Automating the end to end data cataloging process with Waterline Smart Data Catalog significantly reduces the excess cost of redundant data:

- Rationalizes excess data, reducing the cost oft
  - Redundant database software licensing & support
  - Excess server & storage cost
  - Excess data center operating expenses
- Reduces business risk due to:
  - Data stored for government compliances rules and regulations
  - Database performance degradation
  - Questionable data quality

Conclusion

Data rationalization presents a large cost savings opportunity for most IT organizations. By automating the underlying process for inventorying, tagging and curating data and data lineage, redundant data can be identified, rationalized and the savings can be easily harvested. The bottom line is that data redundancy is a much larger hidden cost than most organizations realize. The flip side however is that this is a cost that organizations are already bearing. This means it is also a source to mine for savings and budget. By eliminating the excess storage, database and associated costs of data redundancy, there is a significant vein of gold to mine in cost savings that can be used to fund other data related projects.

Waterline Data

waterlinedata.com

Sales

sales@waterlinedata.com

Technical Support

Visit the Support Center
help@waterlinedata.com

Corporate Headquarters

201 San Antonio Circle Suite 260
Mountain View CA 94040
(650) 946-2104