Governing Big Data and Hadoop

Philip Russom

Senior Research Director for Data Management, TDWI

October 11, 2016

Sponsor

Speakers

Philip Russom
Senior Research Director
for Data Management,
TDWI

Jean-Michel Franco
Director of Product Marketing,
Talend

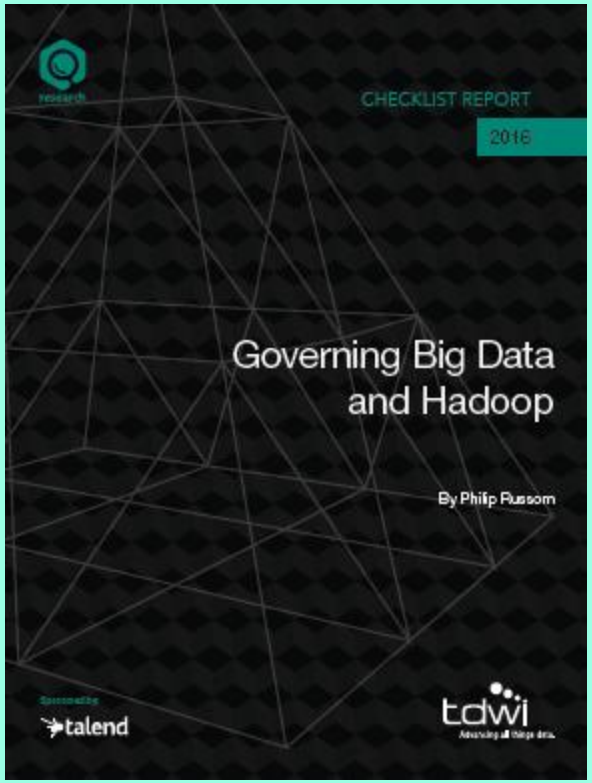
Agenda

- Background
  - *Evolving Data, Management, Business*
  - *Big Data, Hadoop, New Data*
  - *Data Governance (DG)*
- High-Priority Tasks for DG w/Big Data
  1. *Self-service access to big data*
  2. *Data exploration and discovery*
  3. *Metadata for governed big data*
  4. *Simplify DG with integrated tool set*
  5. *Use Hadoop, but govern it carefully*
  6. *DI/DQ Infrastructure to migrate and govern big data*
- Concluding Summary

PLEASE TWEET
@pRussom, #TDWI, @Talend, #BigData, #Hadoop

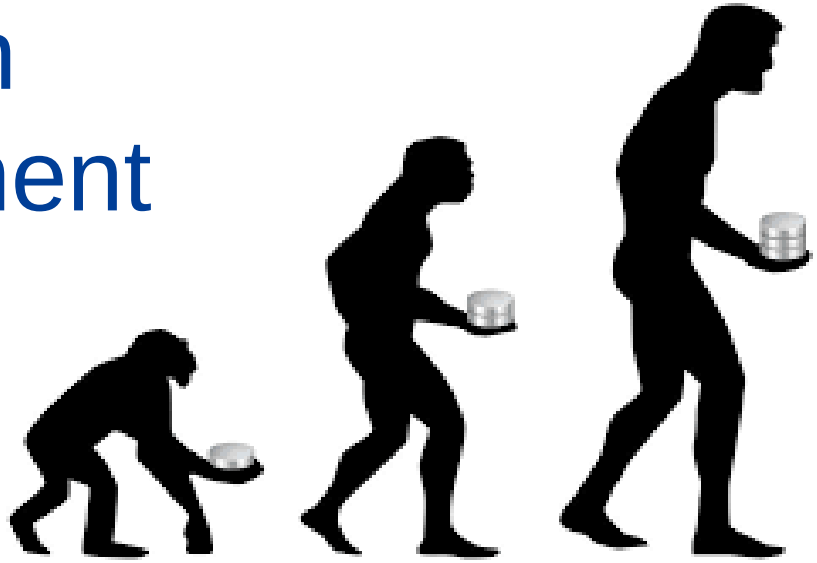
New Checklist Report from TDWI on
Governing Big Data and Hadoop

- The report discusses challenges and solutions for governing Hadoop, big data, and other forms of new data.
- In this webinar, we'll discuss some of the report's findings.
- Stay tuned, to learn how to get a free copy of the whole report.



BACKGROUND

Ongoing Changes in Data & its Management



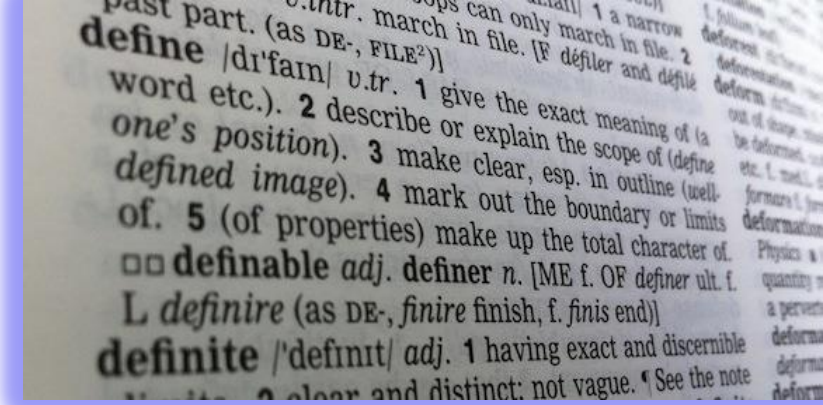
- Data is evolving
- Data management is evolving
- Business use & leverage of data is evolving
- Yet, data must still be governed…

WHAT'S THE REAL POINT?

New Data for New Analytic Insights

- Enable analytics that are new to you
  - *Logistics optimization; Sentiment analysis; Real-time business monitoring*
  - *Patient outcomes in healthcare; Predictive maintenance in manufacturing; Precision farming in agriculture*
  - *Sensor data and other machine data; Streaming data; Human language text*
- Breathe new life into older analytics
  - *Extend existing customer views with additional insights drawn from new big data*
  - *Enlarge data samples of existing analytic applications for fraud, risk, customer segmentation, products of affinity*
  - *Modernize data warehouses, reports, analyses*

DEFINING TWO SIDES OF Data Governance



- Business Compliance for Regulations, Data privacy, and Security
  - *Reducing risks by creating policies for data's access and use*
  - *Identifying sensitive data and guarding it accordingly*
  - *Certifying new big data, so it complies, like other datasets do*
- Technical Standards for Data and Data Management Solutions
  - *Data quality metrics, for both traditional and new big data*
  - *Preferred data modeling techniques*
  - *Preferred interfaces and methods for integration*
  - *Development standards for solution design and dev methods*
  - *Peer review, quality assurance, deployment for DM solutions*

NUMBER ONE

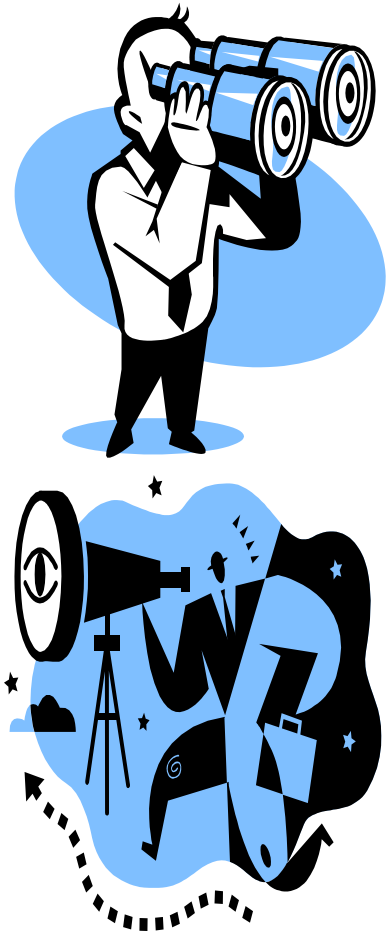
Give people **Self-Service Access** to big data stores, but controlled via data governance.

- Biz & tech people want to work with new data
  - *Business people know that new big data can improve both analytics and operations*
  - *Technology staff need to give a wide range of users access to the new data*
- Many user types want self-service data access
  - *Including data scientists and data analysts, plus some business analysts and managers*
  - *They want to work autonomously and quickly, without waiting for others to create datasets*
- To govern self-service access, tools need:
  - *Security, data protection, auditing via operational metadata, analytic sandboxing for sensitive data*

NUMBER TWO

Integrate new data to enable broad but governed **Data Exploration and Discovery**.

- New big data must be profiled before using
  - *Technical structure, quality, data types*
  - *Business value and compliance issues*
- The point of data exploration is **discovery**
  - *Previously unknown facts about partners, products, customers, operations, financials, logistics…*
  - *New analytic insights from new data*
- Data study and exploration must be governed
  - *Existing DG policies will apply to most new big data*
  - *Some policies may need to be extended, revised*
- Some new big data is inherently sensitive
  - *RE: customers, financials, health, employees, legal*



NUMBER THREE

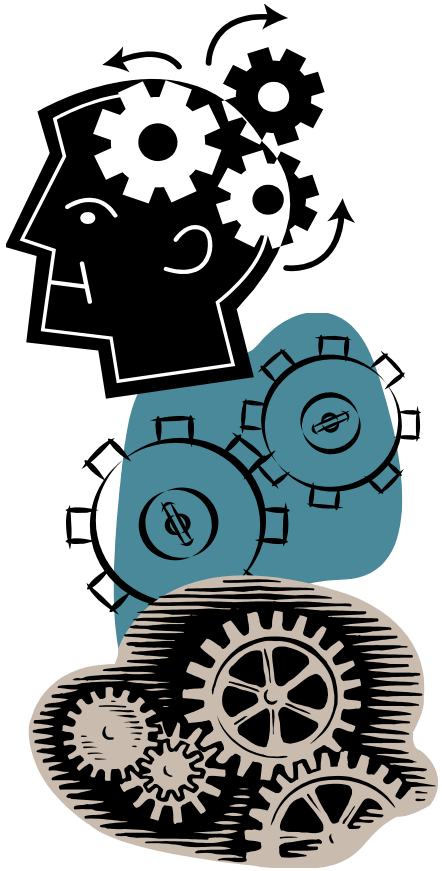
Develop **Metadata for Big Data**, to get the fullest use and governance from it.

- Some new big data lacks obvious metadata
  - *You cannot access source & extract it*
  - *Data has structure, but not described anywhere*
  - *Files may lack header describing schema*
  - *New tools can automatically infer metadata*
- You need all three forms of metadata
  - *Technical – automated software access*
  - *Business – biz-friendly for self-service data access*
  - *Operational – records access events for DG*
- Metadata helps automate data governance
  - *Inventory of data to be governed*
  - *Operational metadata tracks lineage & access*

NUMBER FOUR

Select a platform of **Integrated Data Mgt Tools** for simplified governance via modern solution designs.

- Consider an integrated tool platform
  - *Tools for integration, quality, profiling, metadata, event processing, master data, federation*
  - *New stuff, too: self-service data access, data prep, ad hoc exploration, big data formats & platforms*
  - *All integrated for interoperability in development and during production runs*
- An integrated tool platform supports DG goals:
  - *One tool (instead of several) is easier to govern*
  - *Consistent dev standards and data quality*
  - *Metadata, shared broadly for complete views*
  - *Modern solution designs: in a single auditable data flow, multiple tool types are integrated*



NUMBER FIVE

Consider **Hadoop** for persisting and processing new big data, but beware its governance challenges.

- Hadoop supports desirable use cases
  - *Data landing and staging area for many new big data types and data feed speeds*
  - *Processing for big data, DI, and analytics*
  - *Linear scalability, but at reasonable cost*
  - *Offload your DI, DW, hub, etc.*
- But Hadoop challenges data governance
  - *Data replication*
  - *Metadata management*
  - *Security*
  - *Data store organization*

NUMBER SIX

Provide infrastructure and DG to **Migrate Big Data across complex Data Ecosystems**.

- Multi-platform data architectures are new norm. Most include Hadoop.
  - *Data warehouse environments*
  - *Multi-channel marketing*
  - *Global supply chains*
  - *Web, social, mobile, etc.*
- Requires substantial data integration and quality infrastructure for:
  - *Data movement and migration among platforms and ecosystems*
  - *Broad connectivity and access*
  - *Data lineage, audit, cataloging*
  - *Guided migration, via metadata bridges and interfaces, for automation, standards, and governance for migrations common in today's ecosystems*

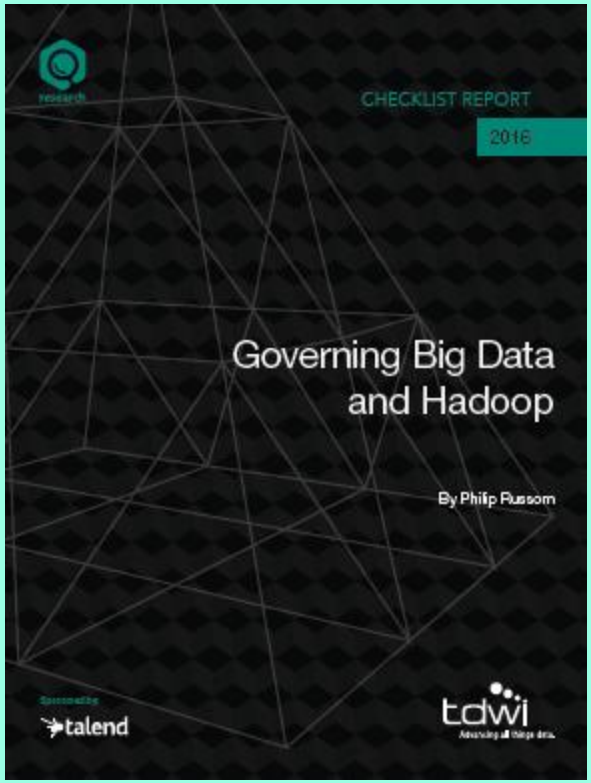
CONCLUDING SUMMARY

Governing Big Data and Hadoop

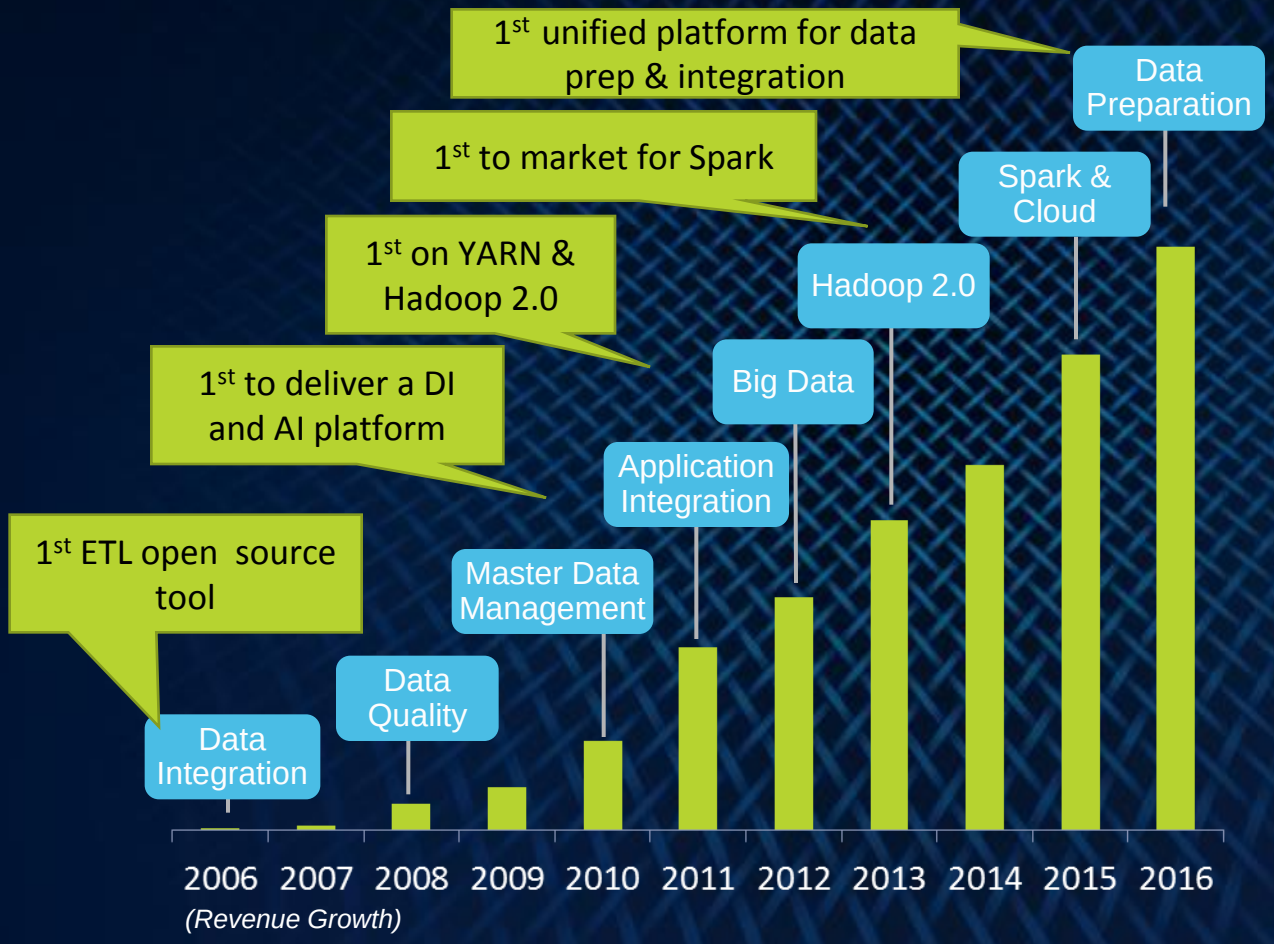
1. Self-service access to big data
2. Data exploration and discovery
3. Metadata for governed big data
4. Simplify DG with integrated tool set
5. Use Hadoop, but govern it carefully
6. DI/DQ Infrastructure to migrate and govern big data

New Checklist Report from TDWI on
Governing Big Data and Hadoop

- The report discusses challenges and solutions for governing Hadoop, big data, and other forms of new data.
- Download the free report in a PDF file at: **www.tdwi.org/checklists**



Talend enables the data-driven enterprise



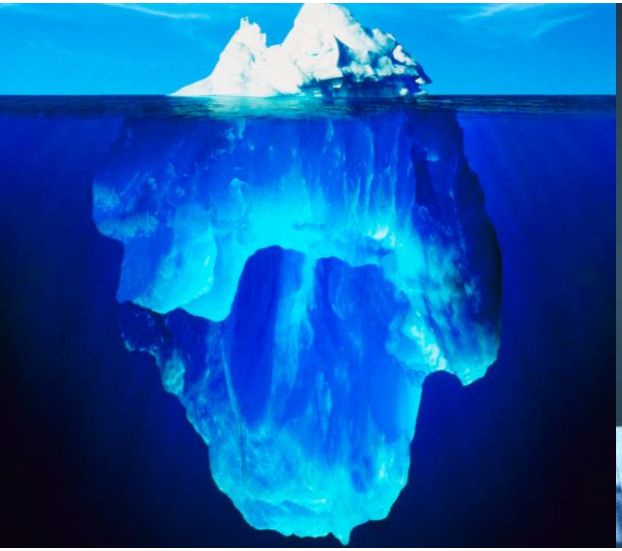
Key Facts

- NASDAQ: TLND (2016)
- $100M Revenue (Projected)
- Leader in Gartner MQ for DI
- 600+ employees worldwide
- 16 countries
- 1300+ customers
- 2M+ open source downloads

talend | Data Fabric

A Modern Big Data and Cloud Integration Platform

Product Investment Themes

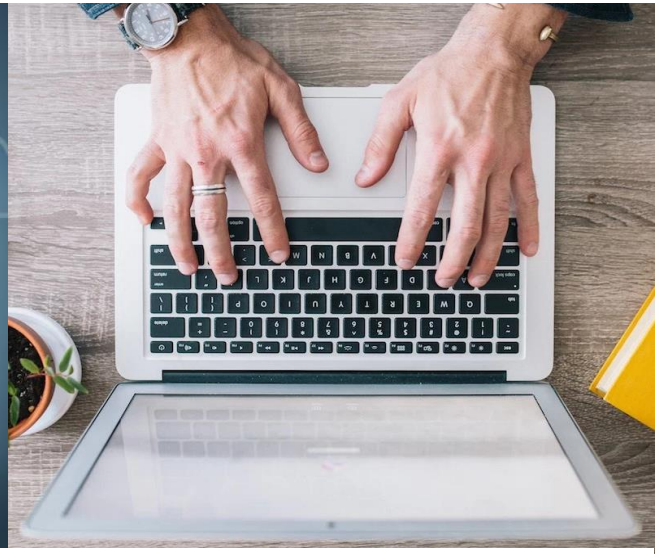


Big Data

Hadoop becomes the dominant data processing platform

Cloud

Cloud overtakes on-premises investments



Self-service

Self-service will be a top IT priority for the next decade

Data Governance is Critical to All Areas

talend

The five pillars for managing Metadata with Talend

Deliver Metadata By Design

Manage and monitor information chains beyond Hadoop

Synchronize Metadata accross data platforms

Democratize the Data Lake with superior accessibility

Tackle Hadoop Governance Challenges

(*) Apache Atlas integration planned in Talend Winter 2017

talend

Redefining patient experiences

Health assistants provide guidance and personalized services to their customers based on individual needs, including data on their lifestyle and health condition, and these are optimized to orient patients to the right care.

75% reduction of efforts and time needed to onboard a new customer

ACCOLADE

talend

Questions?

Thank You to Our Sponsor

Contacting Speakers

- If you have further questions or comments:

Philip Russom
prussom@tdwi.org

Jean-Michel Franco, Talend
jfranco@talend.com