Modernizing the Operational Data Store with Hadoop

Philip Russom

TDWI Research Director for Data Management

January 29, 2014

Thanks to our sponsors

cloudera®

Speakers

Philip Russom
TDWI Research Director,
Data Management

TJ Laher
Product Marketing
Manager,
Cloudera

Shawn James
Director, Big Data
Business Development,
Talend

Agenda

PLEASE TWEET

@pRussom, #TDWI, #Hadoop, #BigData, #DataManagement, #Analytics

- Operational Data Stores (ODSs)
  - *In use for decades, but with new uses today*
  - *Handling big data, analytics, scalable data integration, archiving, ent data hubs…*
  - *In DWs and elsewhere*
- ODSs need modernization
  - *To support new uses, new data, new architectures*
- Hadoop has many uses
  - *Imagine Hadoop as a preferred platform for ODSs*
  - *Scalable, cost effective, flexible, agile, modern*
- Recommendations
  - *Make room for Hadoop, including ODSs on Hadoop*

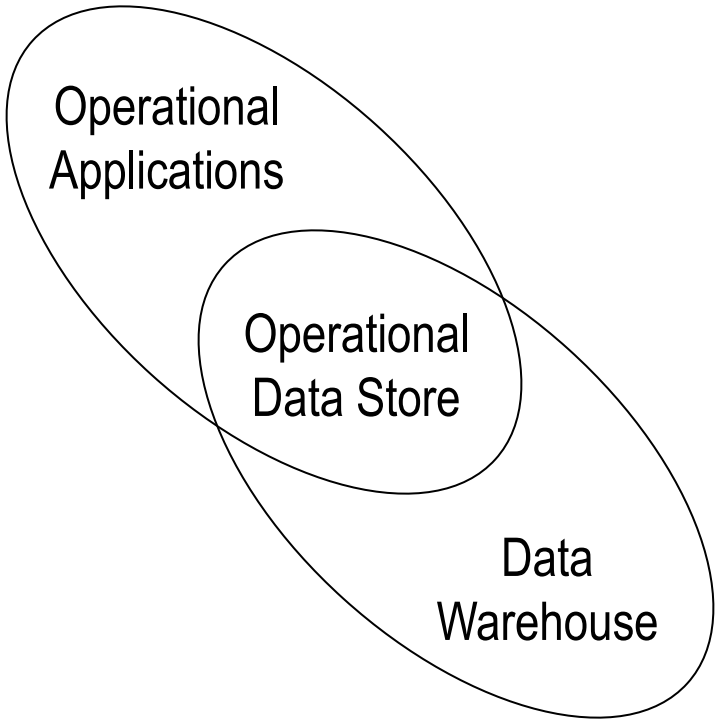
DEFINITION
Operational Data Store

- "The ODS is a basis for doing integrated operational processing, and, in turn, it feeds the data warehouse."
- "The ODS is a separate architectural entity from the data warehouse."
- "An ODS is an architectural construct that is
  - *subject oriented,*
  - *integrated [i.e., aggregated data],*
  - *volatile [updated regularly],*
  - *current valued [little or no archived data],*
  - *and contains only corporate detailed data."*
- "Data in the ODS serves the operational community and as such is kept at a detailed level."
  - *From "Building the Operational Data Store," Bill Inmon et alia*

More on the Operational Data Store

- It's a database
  - *It's a collection of data, designed by users*
  - *Usually running on a relational database management system (RDBMS); but some ODSs are file-based*
  - *Like any database, an ODS can take many forms*
- "Just enough structure"
  - *Simple data models, often just records in a table, few tables/keys*
  - *Data is usually "raw," typically original detailed source data or lightly transformed for standardization*
- Assumption: ODS data will be repurposed
  - *So it's best to keep the original schema, and transform data into new schema, when needed for new analytic applications, etc.*
- Used many ways, in many data architectures
  - *Data warehousing*
  - *Data from or for operational applications*

ODS Use Cases Today



- In DW, BI, DI, analytics
  - *ODSs act as domain specific databases (similar to marts), data landing areas, staging for ETL and other DI processing, archives of source data, real-time buffer*
- In operational applications
  - *Customer masters, in many industries, for master data management, CRM, SFA, lookups, integrating data across customer facing apps, etc.*
  - *Call detail records (CDRs) in telco*
  - *Transaction records in financials*
  - *B2B transactions in supply-chain oriented industries*
- The two above cases via one ODS
  - *The two architectures may overlap via one ODS*

The Evolving ODS and its Uses

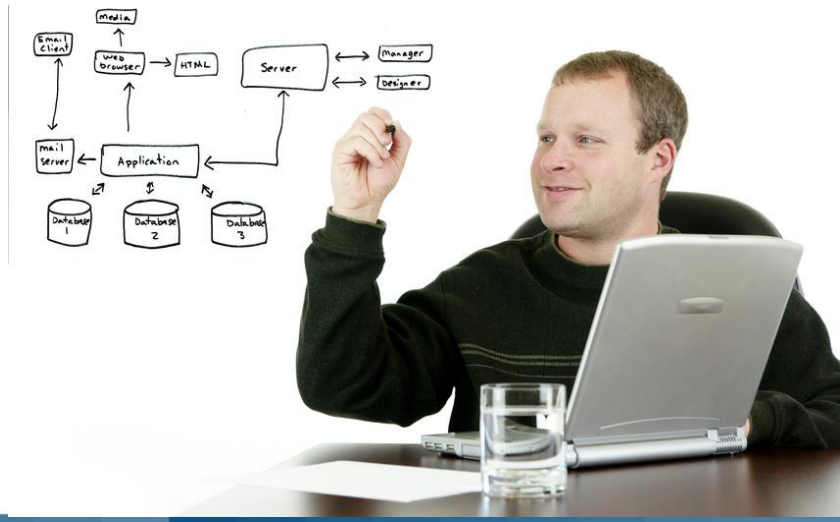
- Driving change in ODS designs and uses
  - *Big data – unlimited scale at a limited price*
  - *Advanced analytics – beyond OLAP to mining, statistics,*
  - *Enable data exploration and discovery*
    - More data in fewer places; less sampling; explore then analyze & visualize
  - *Data integration, landing, staging*
    - Storage for incoming data, before repurposing it; but at greater scale
    - ETL/ELT and analytic processing pushed down into the ODS
  - *Live data archiving – online & queryable for many user types*
- An Important Goal – Multi-purpose ODSs
  - *All the above uses and more via one powerful ODS or a short list of integrated ODSs*
  - *ODS as a consolidation strategy*
  - *AKA: Enterprise data hub*
- Why modernize your ODSs?
  - *To leverage new big data and enable new apps and business uses*
  - *Modernize your EDW and/or other enterprise data*
  - *Work toward the enterprise data hub and/or other consolidation plans*

Using Hadoop as an ODS Platform

- Hadoop is massively scalable
  - *Terabytes and petabytes*
- Hadoop is cost effective
  - *Less than large relational configurations*
  - *Runs well on commodity hardware*
  - *Open-source with affordable maintenance*
- Handles wide range of data types
  - *Both old & new; both structured & not*
- Interoperability and integration via standard interfaces
  - *Many of your existing tools support Hadoop; more tools coming*
- Built for file-based data
  - *Much of the data entering an ODS arrives in files*
  - *Much new data is file based, from sensors, machines, social, etc.*
- Track record of supporting multiple apps & uses
  - *Hadoop 2 & YARN make Hadoop even better multi-user system*

Designing New ODSs

- Choice of platform(s) is key
  - *Probably a mix of relational databases and Hadoop*
    - Plus, file systems, NoSQL, storage subsystems
  - *Data integration platforms with multiple tools*
    - ETL/ELT, federation, services, data quality
- Architecting Multi-Use ODSs
  - *Similar to shared and conformed data modeling*
    - But with simpler schema
    - Or no-schema or new schema
  - *Isolate data & workloads, as in any multi-use system*



Working Toward Modern ODSs

Moving ODS Data to Hadoop

- Collocation
  - *Take control of several ODSs, by moving their data to a central platform*
- Consolidation
  - *Merge multiple ODSs into one or fewer ODSs*
- Migration
  - *This is the larger process of collocating, consolidating, and improving datasets*
- The Fork Lift
  - *Where data moved from one platform to another works well on the new platform with little or no alteration of data*
  - *Users report that ODS data "forklifts" well to Hadoop, typically for use with Hive, HBase, MapReduce, Pig, etc.*

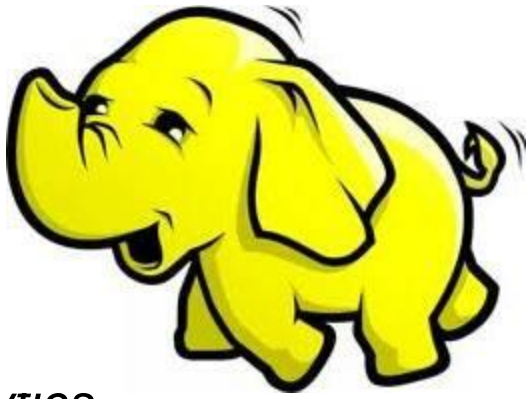
A COMMON ARCHITECTURAL EVOLUTION FOR BIG DATA

Hadoop integrated with a Relational DBMS

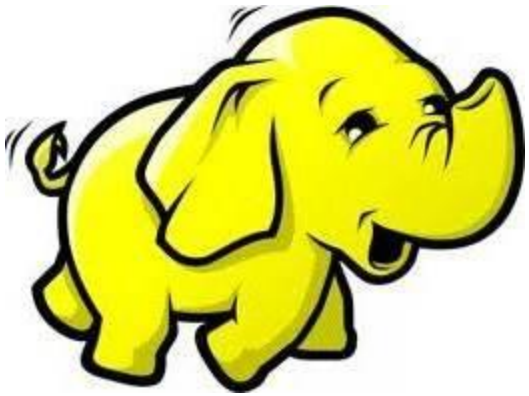
- The strengths of one balance the weaknesses of the other
- A Relational DBMS is good at:
  - *Metadata management*
  - *Complex query optimization*
  - *Query federation*
  - *Table joins, views, keys, etc.*
  - *Security, including roles, directories*
  - *Much more mature development tools*
- HDFS & other Hadoop tools are good at:
  - *Massive scalability*
  - *Lower cost than most DW platforms & analytic DBMSs*
  - *Multi-structured data & no-schema data*
  - *Some ETL functions; late binding; custom code for analytics*
  - *More examples on next slide…*

Data Warehouse Architectures are Evolving to integrate Hadoop

- Some organizations are using Hadoop in multiple areas in their DW architectures
  - *Extension of DW storage*
  - *Operational data stores (ODSs)*
  - *Data staging*
  - *ETL and ELT*
  - *"Archive" of detailed source data, for analytics*
  - *Advanced algorithmic analytics, processed on Hadoop*
  - *Data exploration, discovery, and visualization*
- Even when the above migrate to Hadoop, the core DW still provides data for the majority of BI deliverables:
  - *Standard reports, dashboards, performance management, OLAP*



It's not just Data Warehouses. Hadoop has Other Enterprise Uses.

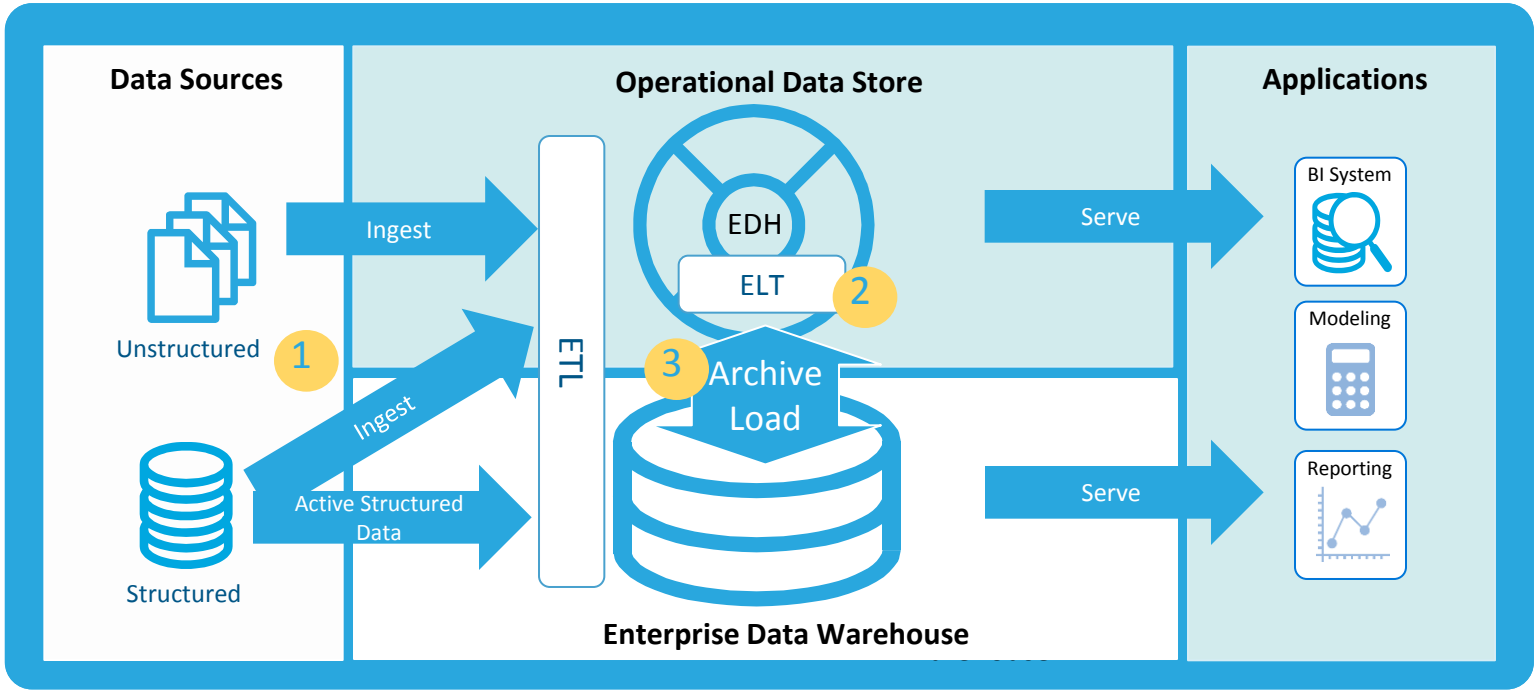


- Data archiving
  - *Most data archives are old and useless*
  - *Hadoop can enable a modern "live archive" that's massively scalable and accessible at any moment by any user*
- Content management
  - *Most "content" is file-based and requires massively scalable search*
  - *Hadoop excels with those, plus adds broad analytics for content*
- Storage as a shared enterprise asset
  - *IT provides SAN/NAS*
  - *Why not Hadoop, too?*

Triggers for ODSs on Hadoop

- To modernize an existing data warehouse
  - *Migrate ODSs off DW platform to Hadoop*
  - *Free up DW capacity instead of buying more*
- To capture new data and big data
  - *An ODS on Hadoop is ideal for log data, sensor data, machine data, device data*
- To consolidate data from multiple platforms
  - *Consolidate ODSs for better exploration, governance, analytics*
  - *Consolidate ODSs on Hadoop for low-cost admin & processing*
- To deploy an Enterprise Data Hub (EDH) on Hadoop
  - *EDH designs vary, but most users create an EDH that resembles an ODS or a series of ODSs*
  - *Users are trending to fewer, larger, multi-use ODSs, as in EDH*

The Modern Architecture



1) Ingest More Data	2) Optimize Data Processing	3) Automated Secure Archive

cloudera®

©2014 Talend Corporation

Main Challenges in the Data Integration Market

BIG DATA

More data, less structure

PRODUCTIVITY

Can't keep up with demand

SKILLS

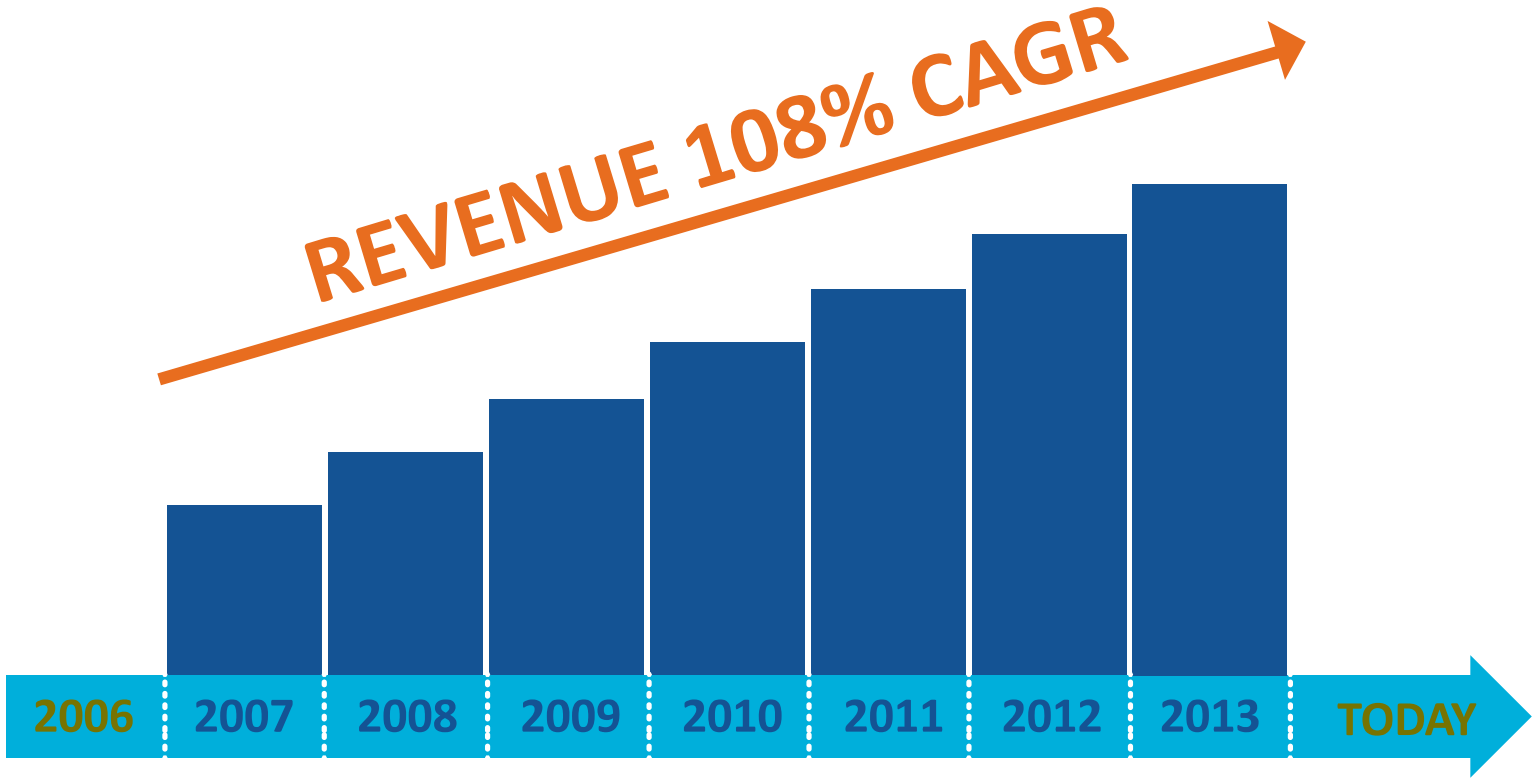
Hard to find talent

COST

Expensive solutions

©2014 Talend Corporation

Introducing Talend, a Disruptive Leader



Creation

talend

"Visionary"

Gartner

"Leader"

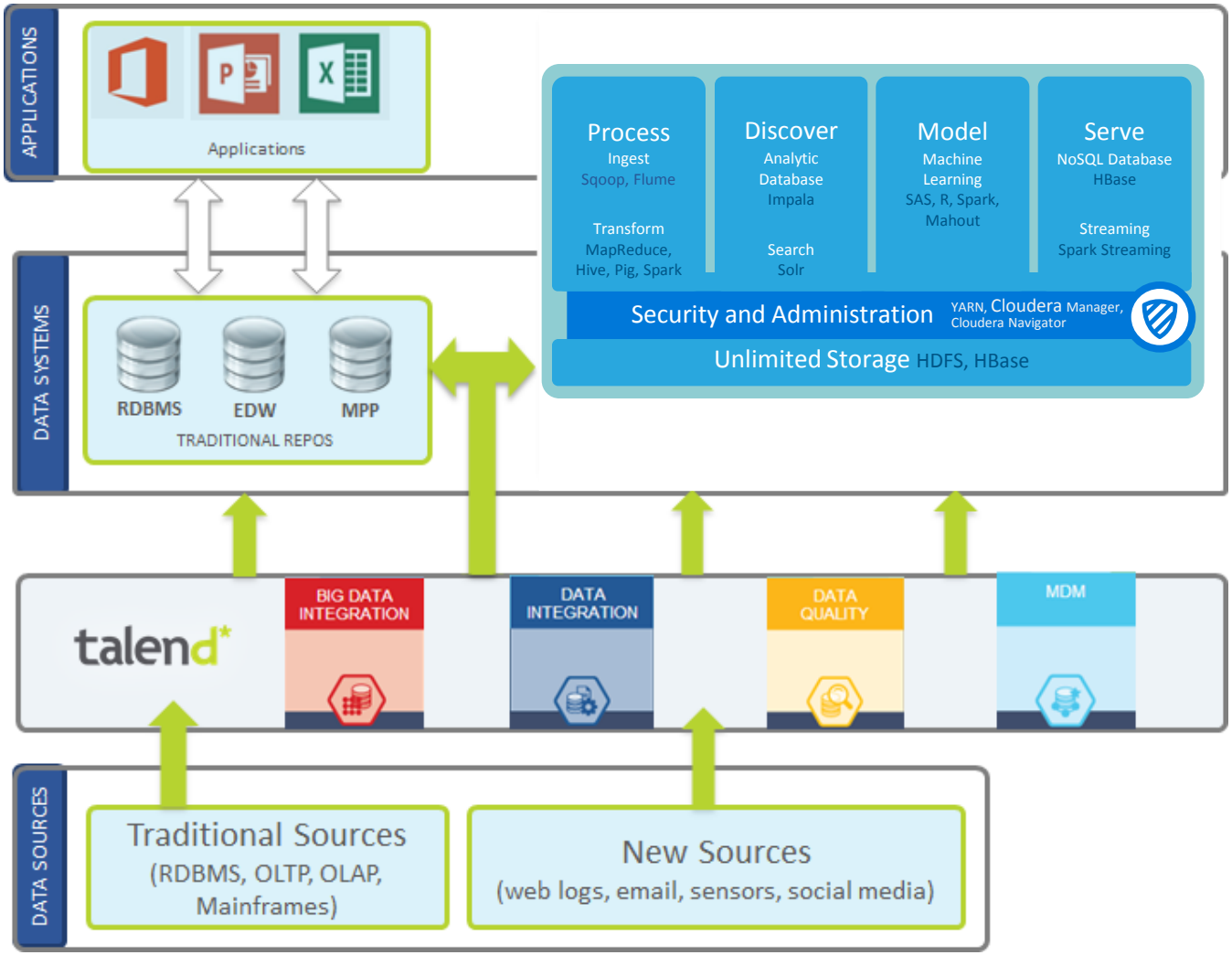
FORRESTER®

Customers

Allianz Global Investors
AXA
KAISER PERMANENTE
orange
ING
Virgin mobile
The Weather Channel
citi

©2014 Talend Corporation

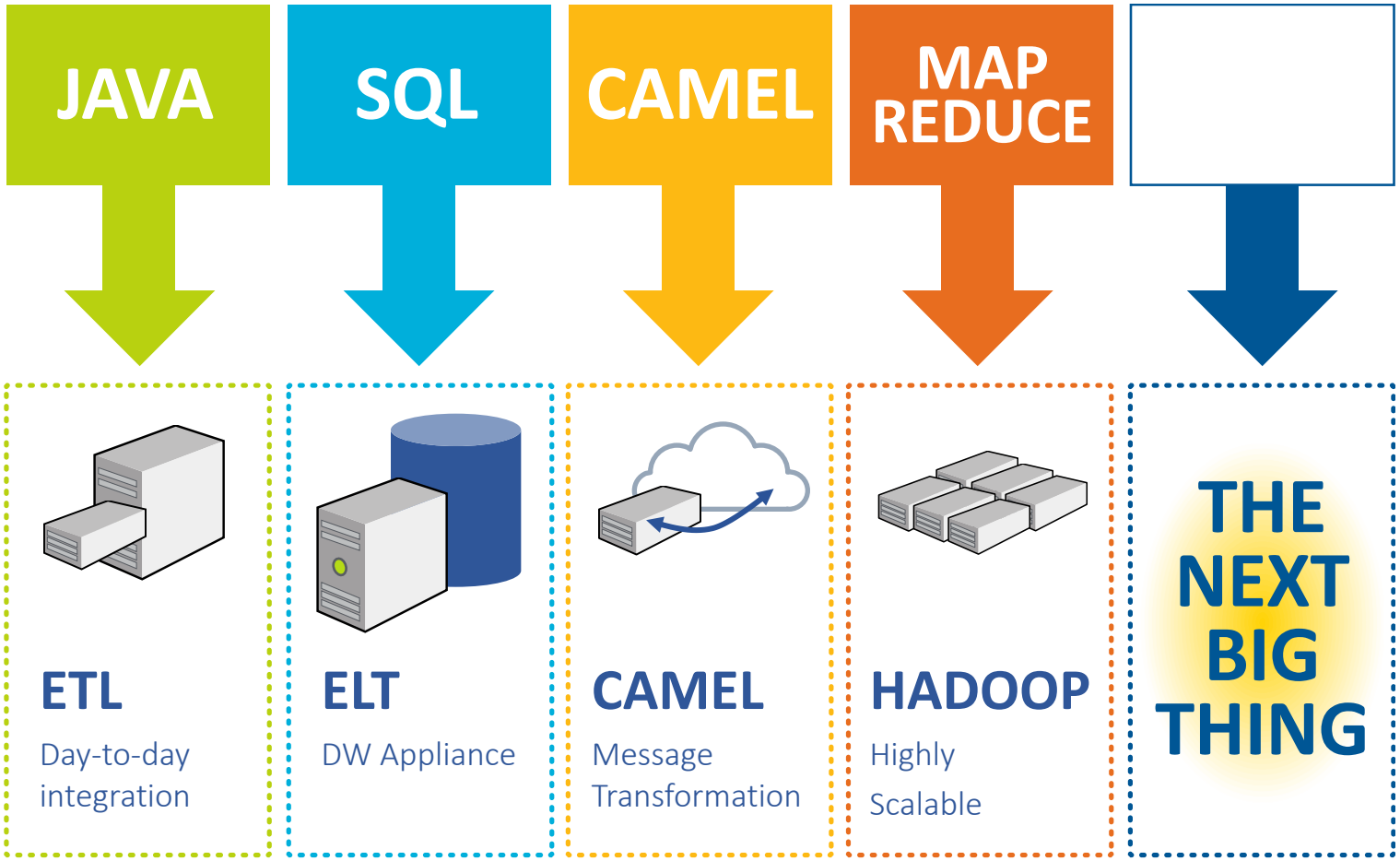
Cloudera / Talend Integration



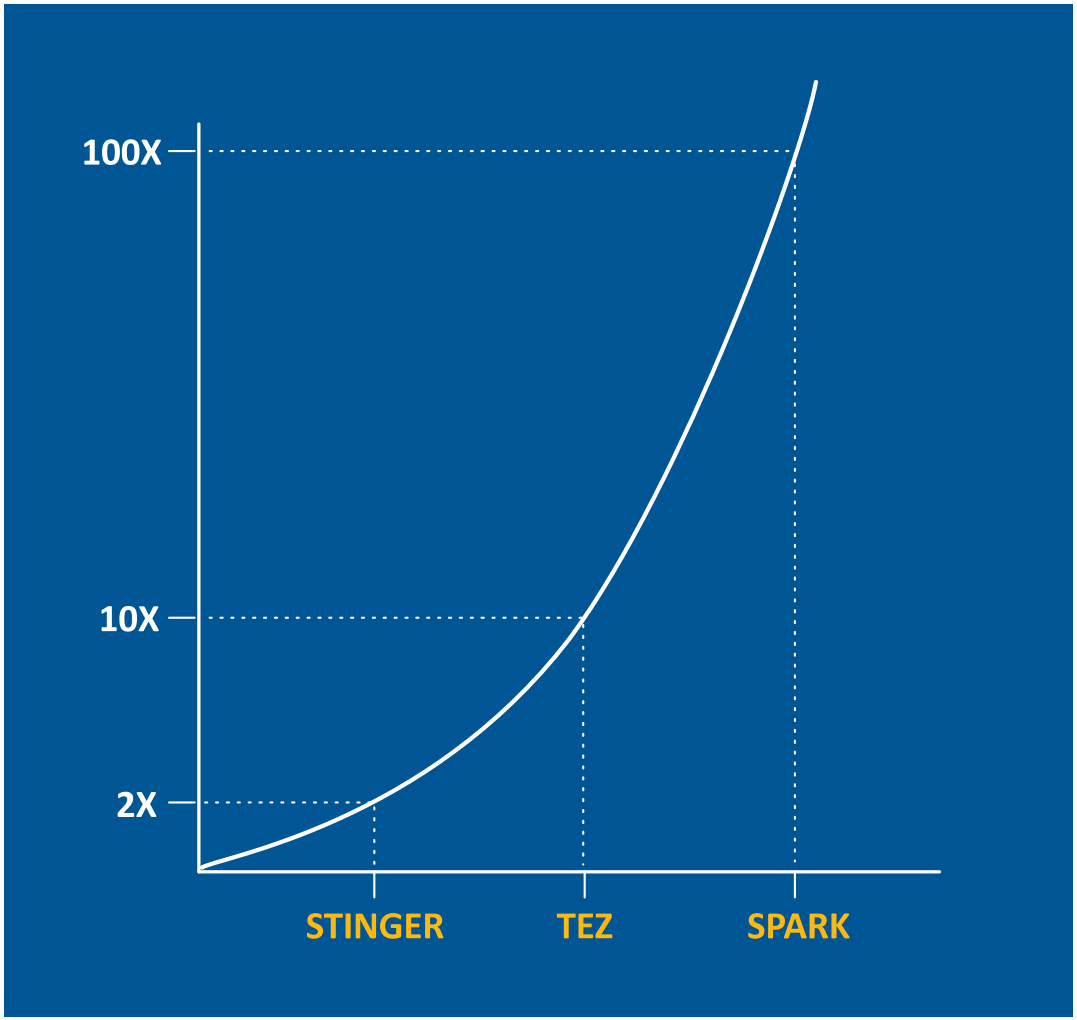
Talend Big Data

Easiest and Most Powerful Integration Solution for Big Data

Future-Proof Architecture With Native Code Gen



The Performance Benefits of Native



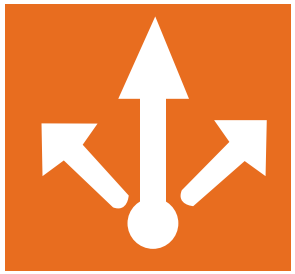
ONLY TALEND RUNS NATIVE ON HADOOP

- 1st on MapReduce
- 1st on YARN
- 1st on Spark (preview)
- 1st on Storm (preview)

©2014 Talend Corporation

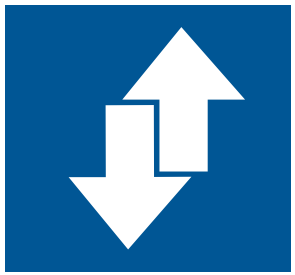
Main Challenges in the Data Market

SCALABLE



AGILE

EASY



LOWEST TCO

©2014 Talend Corporation

1,800 Leading Brands Use Talend

FINANCE & INSURANCE

SERVICES

MANUFACTURING & RETAIL

PUBLIC SECTOR & EDUCATION

©2014 Talend Corporation

Questions?

Contact Information

- If you have further questions or comments:

Philip Russom
prussom@tdwi.org

Shawn James
sjames@talend.com

TJ Laher
tlaher@cloudera.com