HORTONWORKS®

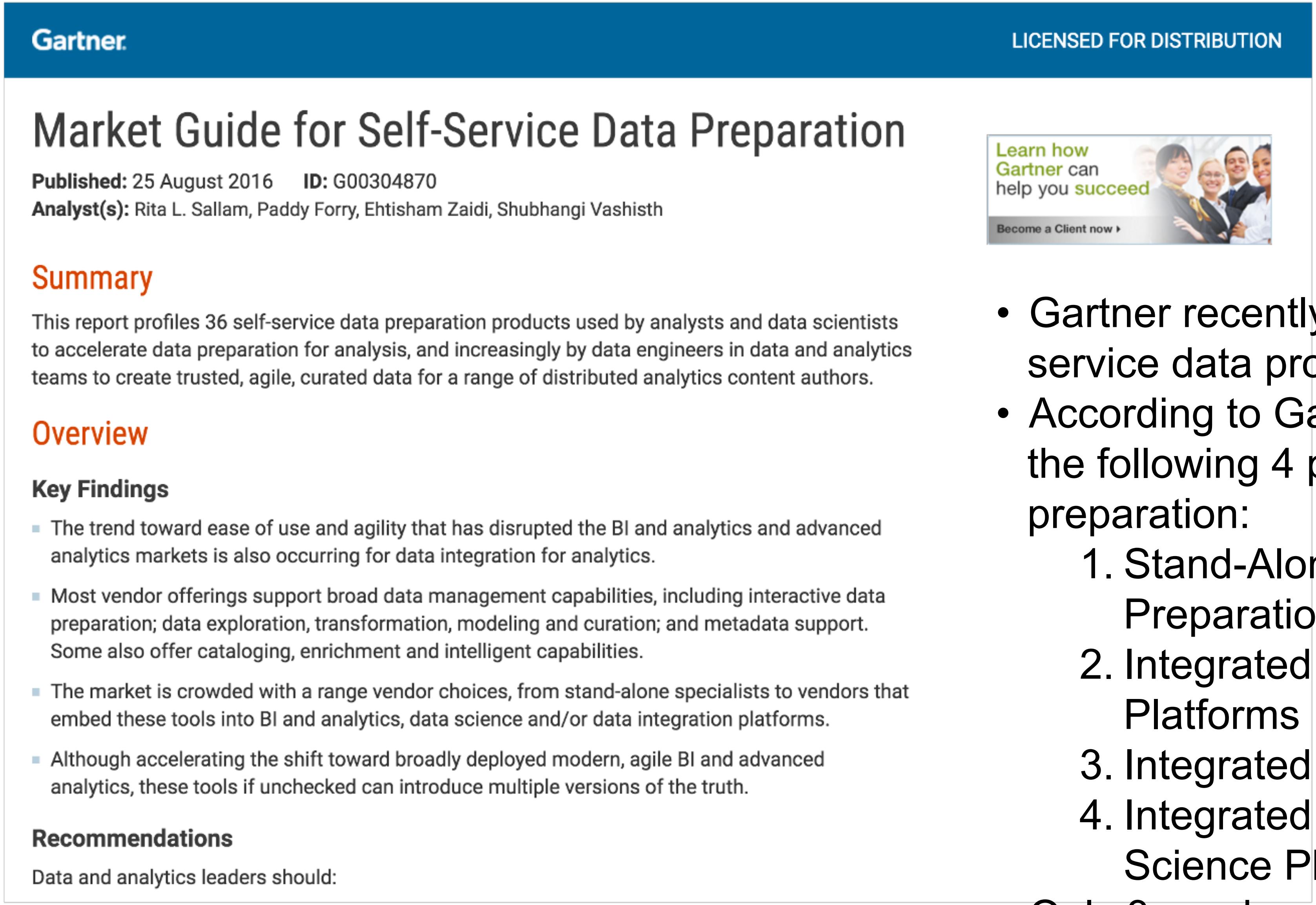
pentaho®
A Hitachi Group Company

Pentaho
Technical Overview

Max Felber
Solution Engineer
mfelber@pentaho.com

September 22, 2016

Industry Leader in Self-Service Big Data Preparation



Gartner

LICENSED FOR DISTRIBUTION

Market Guide for Self-Service Data Preparation

Published: 25 August 2016 **ID:** G00304870

Analyst(s): Rita L. Sallam, Paddy Forry, Ehtisham Zaidi, Shubhangi Vashisth

Summary

This report profiles 36 self-service data preparation products used by analysts and data scientists to accelerate data preparation for analysis, and increasingly by data engineers in data and analytics teams to create trusted, agile, curated data for a range of distributed analytics content authors.

Overview

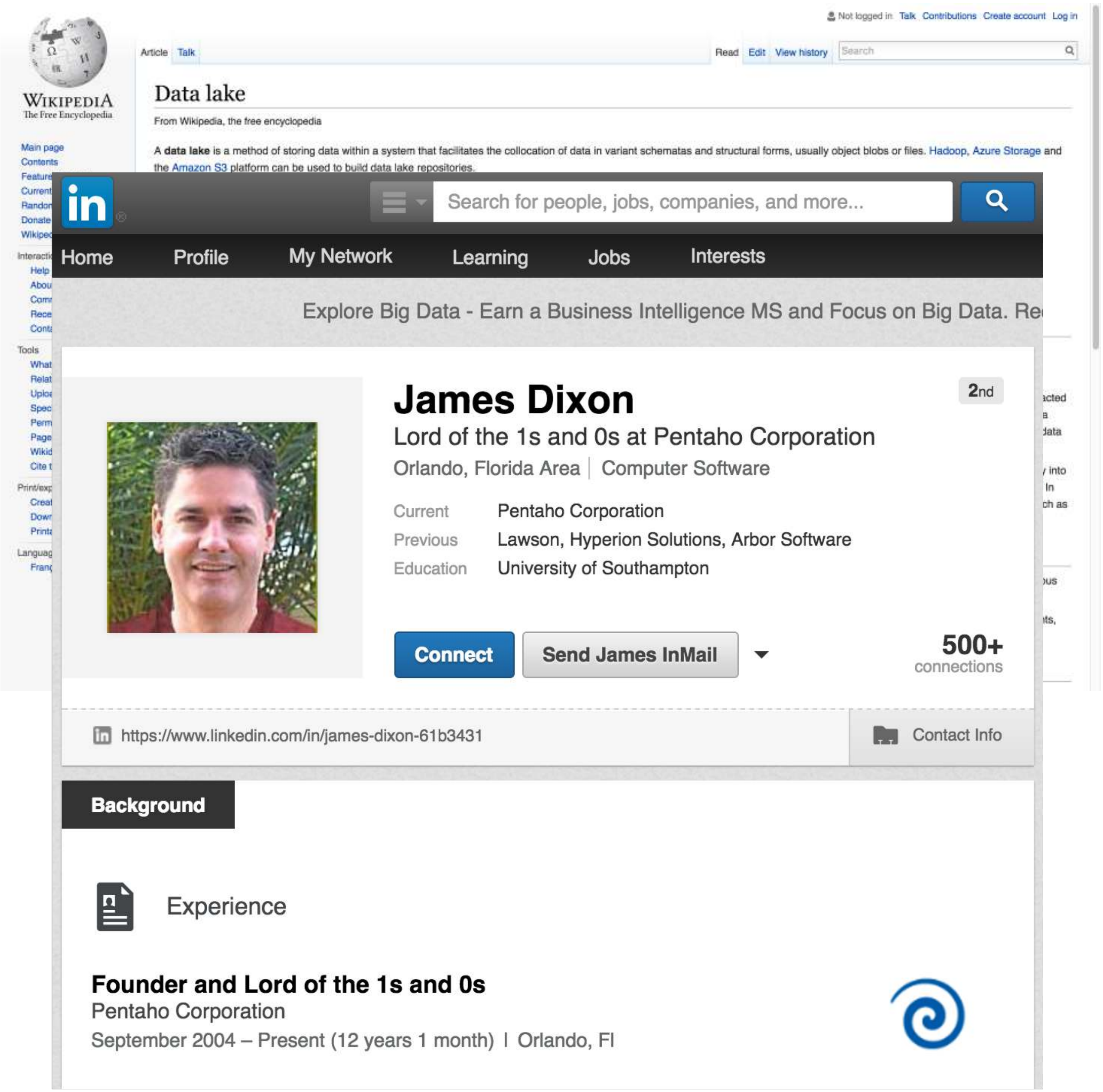
Key Findings

- The trend toward ease of use and agility that has disrupted the BI and analytics and advanced analytics markets is also occurring for data integration for analytics.
- Most vendor offerings support broad data management capabilities, including interactive data preparation; data exploration, transformation, modeling and curation; and metadata support. Some also offer cataloging, enrichment and intelligent capabilities.
- The market is crowded with a range vendor choices, from stand-alone specialists to vendors that embed these tools into BI and analytics, data science and/or data integration platforms.
- Although accelerating the shift toward broadly deployed modern, agile BI and advanced analytics, these tools if unchecked can introduce multiple versions of the truth.

Recommendations

Data and analytics leaders should:

- Gartner recently completed a study on 36 self-service data providers [Gartner Report]
- According to Gartner, a vendor should fulfill the following 4 pillars of self-service data preparation:
  1. Stand-Alone Self-Service Data Preparation
  2. Integrated With Existing Data Integration Platforms
  3. Integrated With Modern BI&A Platforms
  4. Integrated With Advanced Analytics/Data Science Platforms
- Only 3 vendors met each pillar: Oracle, IBM and Pentaho



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history

WIKIPEDIA
The Free Encyclopedia

Main page
Contents

Data lake

From Wikipedia, the free encyclopedia

A **data lake** is a method of storing data within a system that facilitates the collocation of data in variant schematas and structural forms, usually object blobs or files. Hadoop, Azure Storage and the Amazon S3 platform can be used to build data lake repositories.

Explore Big Data - Earn a Business Intelligence MS and Focus on Big Data. Re

James Dixon

2nd

Lord of the 1s and 0s at Pentaho Corporation

Orlando, Florida Area | Computer Software

Current	Pentaho Corporation
Previous	Lawson, Hyperion Solutions, Arbor Software
Education	University of Southampton

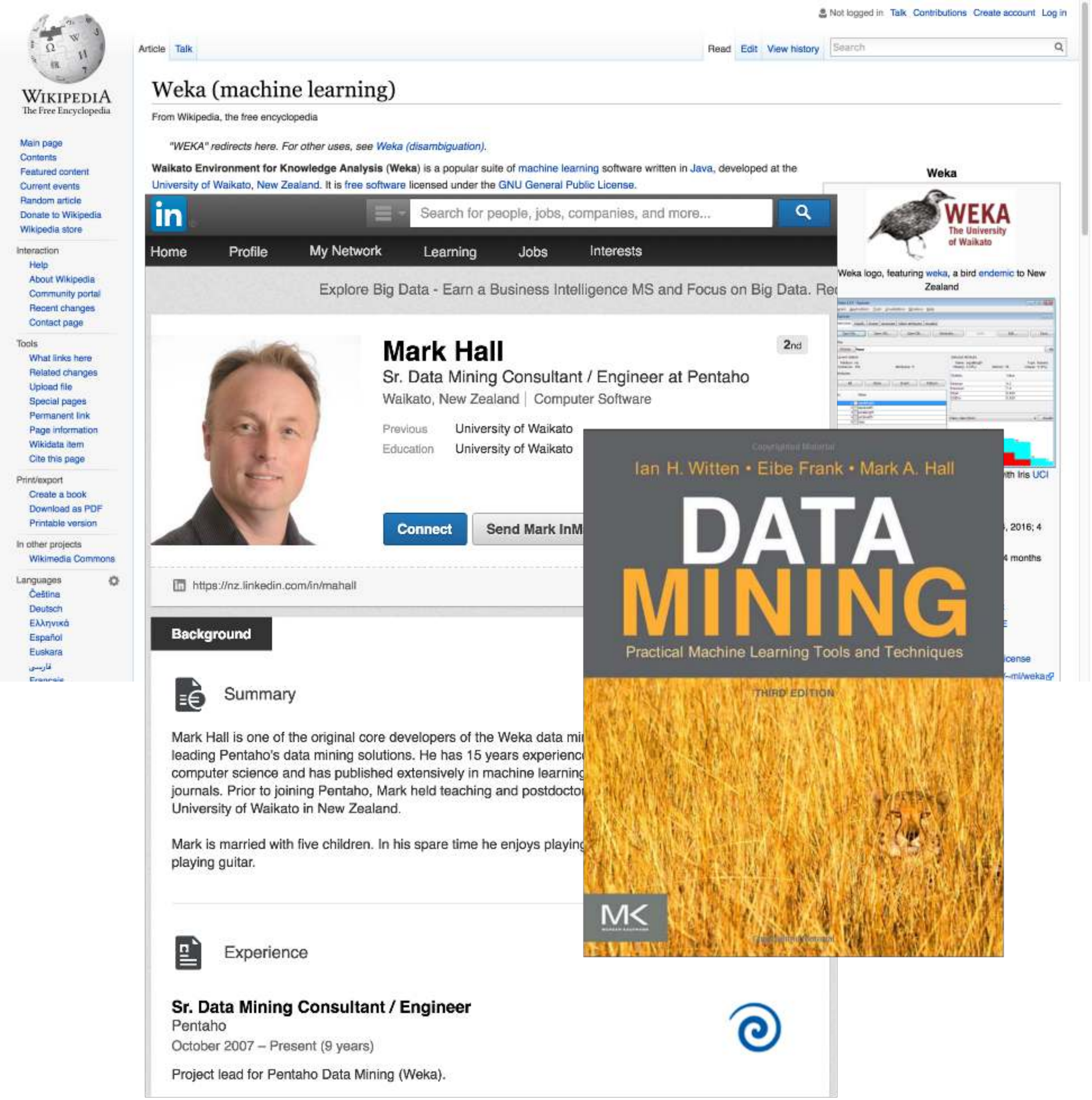
500+
connections

https://www.linkedin.com/in/james-dixon-61b3431

Experience

Founder and Lord of the 1s and 0s
Pentaho Corporation
September 2004 – Present (12 years 1 month) I Orlando, Fl

- How often have you heard the term Data Lake?
- James Dixon, our founder, invented the term



Weka (machine learning)

From Wikipedia, the free encyclopedia

"WEKA" redirects here. For other uses, see Weka (disambiguation).

Waikato Environment for Knowledge Analysis (**Weka**) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Weka

Weka logo, featuring weka, a bird endemic to New Zealand

WEKA
The University of Waikato
Search for people, jobs, companies, and more...
Home
Profile
My Network
Learning
Jobs
Interests

Explore Big Data - Earn a Business Intelligence MS and Focus on Big Data. Rec

Mark Hall

2nd

Sr. Data Mining Consultant / Engineer at Pentaho

Waikato, New Zealand | Computer Software

Previous: University of Waikato

Education: University of Waikato

https://nz.linkedin.com/in/mahall

Background

Summary

Mark Hall is one of the original core developers of the Weka data mi leading Pentaho's data mining solutions. He has 15 years experienc computer science and has published extensively in machine learning journals. Prior to joining Pentaho, Mark held teaching and postdocto University of Waikato in New Zealand.

Mark is married with five children. In his spare time he enjoys playing playing guitar.

Experience

Sr. Data Mining Consultant / Engineer
Pentaho
October 2007 – Present (9 years)

Project lead for Pentaho Data Mining (Weka).

Ian H. Witten • Eibe Frank • Mark A. Hall

DATA MINING

Practical Machine Learning Tools and Techniques

THIRD EDITION

- Maybe you've heard of Weka for machine learning?
- Mark Hall, the man who leads our data mining services not only developed it, he wrote the book on it

So why do customers choose Pentaho?

1. Metadata Injection. By utilizing dynamic vs. static bindings, we can reduce the number of data transformations by 80-90 percent. YouTube Demo by Matt Casters, Chief Architect of Data Integration and Kettle Project Founder at Pentaho
2. Visual MapReduce. Graphically build Hadoop data transformations without coding. This enables you to reduce your Hadoop development time by over 85 percent. Additionally, Pentaho automatically manages the deployment and execution of Hadoop transformations with YARN. YouTube Demo by Doug Moran, Product Manager for Big Data Technologies and Co-Founder of Pentaho
3. Embedded Analytics. “White label” our reports, visualizations and dashboards directly into your web applications. Use Java and REST APIs to access Pentaho data transformations and reports. YouTube Demo by Anthony de Shazor, SVP of Customer Care and Principal Architect
4. Beyond ETL. Pentaho supports Enterprise Information Integration (EII), also known as data federation. No you can create ETL jobs that blend data from structured and big data sources and invoke it via JDBC with Teiid
5. Weka Scoring, Forecasting and R script execution with our Data Science Pack. This package helps Data Scientists reduce data preparation times by 60-80 percent
6. Deep Integration with your Cloudera, Hortonworks or MapR Hadoop ecosystem. Pentaho offers real control over YARN jobs, Spark execution, Oozie, Sqoop and more
7. 2,000 Commercial Customers and 20,000 Production Deployments

Best of Breed vs. Best Platform

VS.

It’s all about Balance. Focus on the Integration and Workflow.

HITACHI

pentaho
A Hitachi Group Company

Pentaho – Part of Hitachi Group Companies

HITACHI
Inspire the Next

Digital Operations, Digital Enterprise, Digital Customer Experience & Smart Infrastructure

Hitachi Data Systems

Turnkey BI and Big Data Solutions
Digital Transformation & Social

pentaho
A Hitachi Data Systems Company

Big Data Integration & Analytics.
Embedded or SaaS model

Big Data – since 2009

Traditional DI & BI
Since 2004

HBASE
Cassandra
mongoDB
elasticsearch
splunk>
CouchDB
relax
VERTICA
An HP Company
HANA

SSO &
Java Spring

DB per
Group/Tenant

Row-level
Multi-tenancy

Object
Multi-tenancy

UI
Multi-tenancy

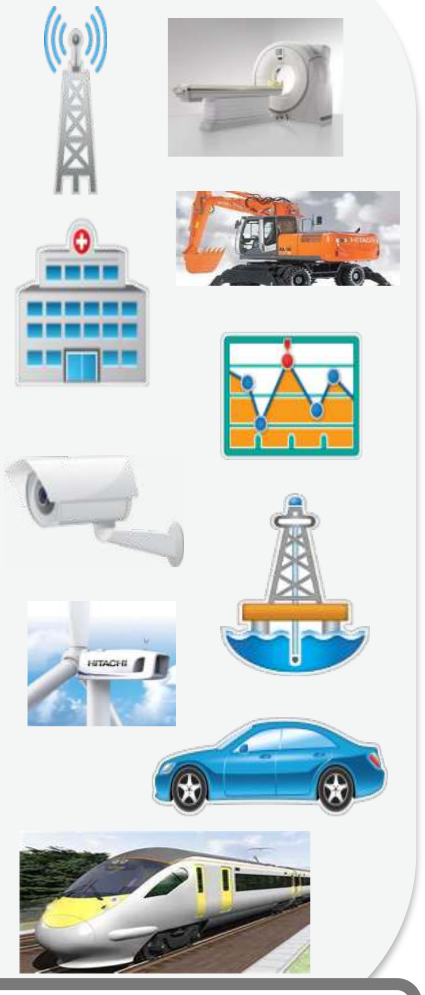
Scale-out
Architecture

Unified
Compute
Platform (UCP)

pentaho
A Hitachi Data Systems Company

Hyper-Scale-out
Platform (HSP)

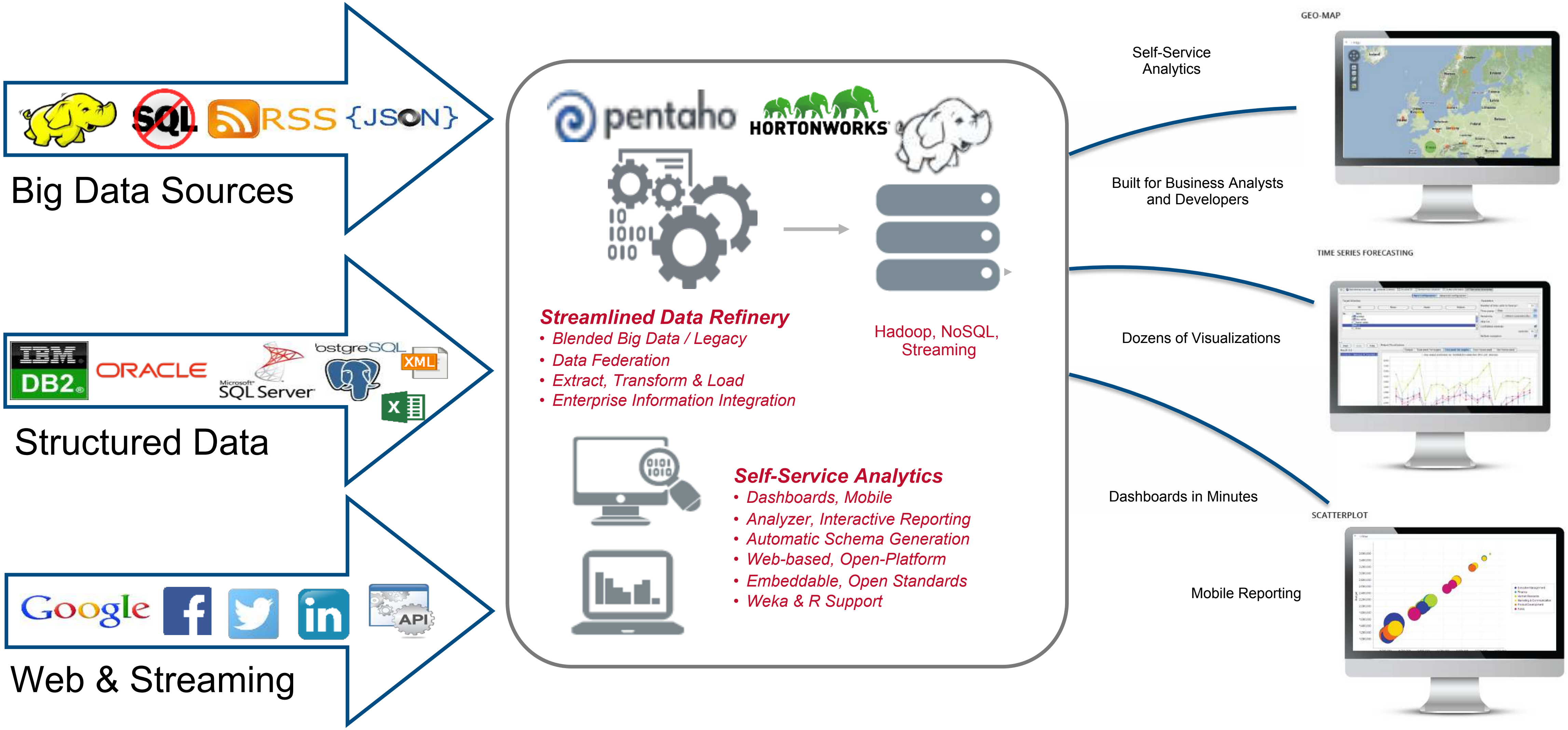
Hortonworks



Hitachi Consulting

We make it happen. Better.

End-to-End Big Data Integration, Embedded Analytics & Dashboards



Big Data Sources

Structured Data

Web & Streaming

Streamlined Data Refinery

- *Blended Big Data / Legacy*
- *Data Federation*
- *Extract, Transform & Load*
- *Enterprise Information Integration*

Hadoop, NoSQL, Streaming

Self-Service Analytics

- *Dashboards, Mobile*
- *Analyzer, Interactive Reporting*
- *Automatic Schema Generation*
- *Web-based, Open-Platform*
- *Embeddable, Open Standards*
- *Weka & R Support*

Self-Service Analytics

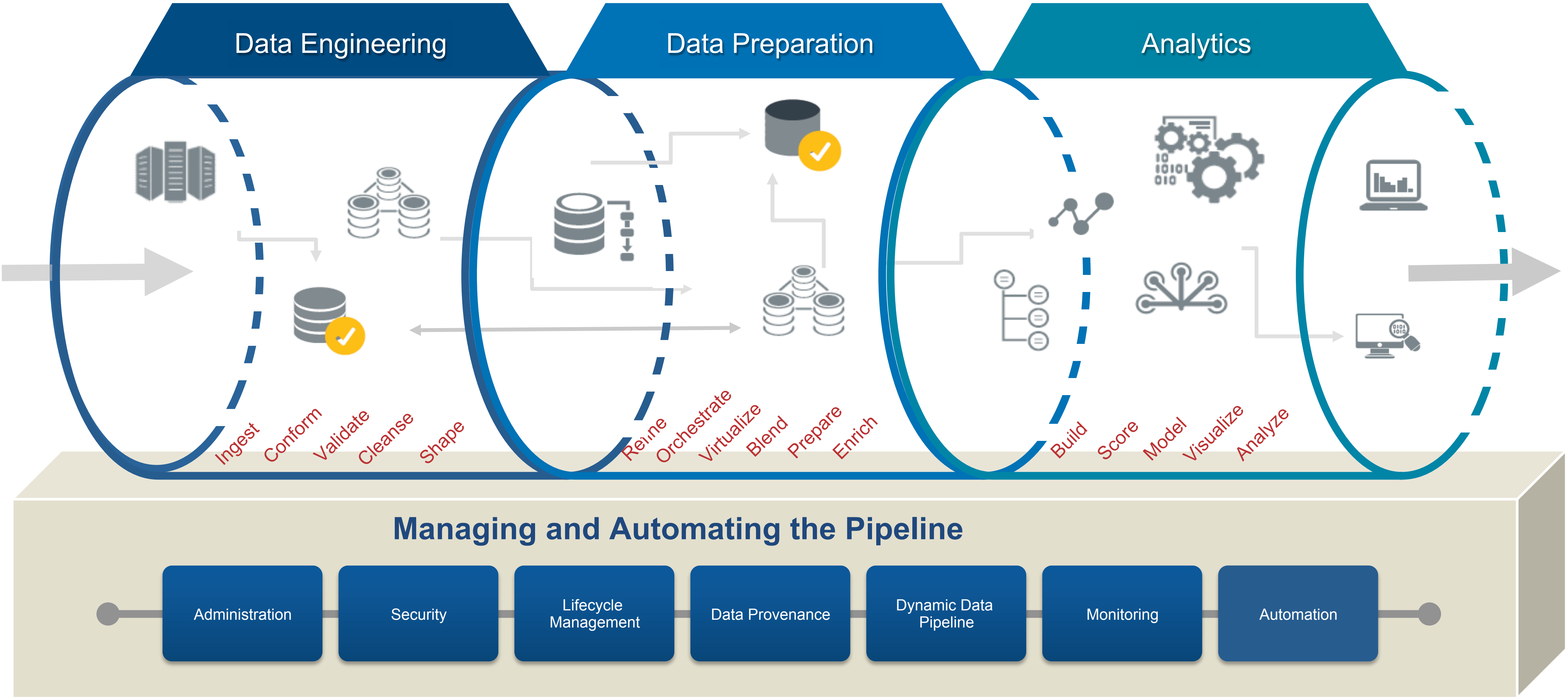
Built for Business Analysts and Developers

Dozens of Visualizations

Dashboards in Minutes

Mobile Reporting

End-to-End Development Solution



Data Engineering

Data Preparation

Analytics

Ingest

Conform

Validate

Cleanse

Shape

Refine

Orchestrate

Virtualize

Blend

Prepare

Enrich

Build

Score

Model

Visualize

Analyze

Managing and Automating the Pipeline

Administration

Security

Lifecycle Management

Data Provenance

Dynamic Data Pipeline

Monitoring

Automation

So why do customers buy Pentaho?

1. Metadata Injection. By utilizing dynamic vs. static bindings, we can reduce the number of data transformations by 80-90 percent. YouTube Demo by Matt Casters, Chief Architect of Data Integration and Kettle Project Founder at Pentaho
2. Visual MapReduce. Graphically build Hadoop data transformations without coding. This enables you to reduce your Hadoop development time by over 85 percent. Additionally, Pentaho automatically manages the deployment and execution of Hadoop transformations with YARN. YouTube Demo by Doug Moran, Product Manager for Big Data Technologies and Co-Founder of Pentaho
3. Embedded Analytics. “White label” our reports, visualizations and dashboards directly into your web applications. YouTube Demo by Anthony de Shazor, SVP of Customer Care and Principal Architect
4. Current Big Data Projects Struggling or Failing
5. Internet of Things (IoT) Initiatives
6. On-Boarding Initiatives (Consolidation, Data Warehousing, SaaS)
7. 360 Degree View (Blending Traditional & Big Data Sources)
8. Predictive Analytics (“R”) & Machine Learning (“Weka”)
9. Embedding (White-Labeling) Reporting & Analytics
10. ERP Migration
11. Cloud Deployments
12. Data Federation (a.k.a. Enterprise Information Integration)
13. Blended Application and Data Layer Integration (SOA, Web Services, etc.)

Thank You