How to Build a Successful Data Lake

May 17, 2016
Before We Begin

• This webinar is being recorded. Later this week, you will receive an email on how to get the recording and slide deck.

• If you have any audio problems, please let us know in the chat window and we’ll try to resolve them quickly.

• If you have any questions during the webinar, please type them in the chat window.
Introducing Our Speakers

Dale Kim
Sr. Director, Industry Solutions
MapR Technologies

Alex Gorelik
Founder and CEO
Waterline Data
How to Build a Successful Data Lake

Dale Kim, Sr. Director, Industry Solutions, MapR Technologies

May 17, 2016
What to Consider for Your Platform

• Broad analytics capabilities
• Interoperability
• Business continuity
• Cost effectiveness
• Multi-tenancy capabilities
Broad Analytics Capabilities

• Human analytics
  – Visualizations – graphs, charts, pictures
  – Obvious insights when presented in the right way

• Algorithmic analytics
  – Heavy computations
  – Finding non-obvious trends and alerting a system or a human
Interoperability

- RDBMSs and Data Warehouses
- NoSQL Databases
- Event Processing Systems
- Enterprise Storage
Business Continuity

• High availability – tolerance for multiple hardware failures in a data center
• Disaster recovery – fast failover to a remote site
• Data recovery – quickly restore from data corruption from user/app errors
Cost Effectiveness

• Any combination of:
  – Lower hardware footprint
  – Lower admin. overhead
  – Higher performance
  – Greater resource sharing
Multi-Tenancy Capabilities
How to Build a Successful Data Lake

Alex Gorelik
Founder and CEO, Waterline Data
Waterline Data Overview

Investors

Alex Gorelik
Founder, CEO
Founded Exeros (IBM) and Acta (SAP), IBM DE, Informatica GM, MSCS Stanford, Columbia BSCS

Oliver Claude
Marketing
VP SAP, VP Informatica, IBM Siebel, Nova Southeastern MS MIS

Jason Chen
Engineering
VP Teradata, Acta, Sybase. USC PhD CS.

Ravi Ramachandran
Sales
CSC Infochimps, AppLabs, Xchanging. Scient-Razorfish. MBA Clark, BS Delhi University.

Mohan Sadashiva
Product
Narus (Boeing), Intel, Synchronoss, Trimble Navigation. MBA Columbia, MSCS Queens University

Advisors

Google

LinkedIn

MapR

Partners

Amazon Web Services

CSC

Infosys

Kaiser Permanente

Trifacta

Syncsort
Production Customers by Industry

**Healthcare**
- Fortune 500
- Healthcare Provider

**Insurance**
- Fortune 500 Health Insurer & Global Insurer

**Government**
- Government Agency in EMEA

**Automotive**
- Leading US Vehicle Remarketing Provider

**Consumer Marketing**
- Leading Market Research Firm in EMEA
Data Lakes Power Data Driven Decision Making
Business Value

Data Swamp
- No Value

Data Warehouse Off-loading
- Cost Savings

Data Puddles
- Limited Scope and Value

Data Lake
- Enterprise Impact
Data Swamps

Raw data

Can’t find or use data

Can’t allow access without protecting sensitive data
I prefer a data warehouse – it’s more predictable.

It takes IT 3 months of data architecture and ETL work to add new data to the data lake.

I can’t get the original data.
Data Puddles: Limited Scope and Value
Data Puddles: Limited Scope and Value

Low variety of data and low adoption
- Focused use case (e.g., fraud detection)
- Fully automated programs (e.g., ETL off-loading)
- Small user community (e.g., data science sand box)

Strong technical skill set requirement
What Makes a Successful Data Lake?

Right Platform + Right Data + Right Interface
Right Platform:

- Volume - Massively scalable
- Variety - Schema on read
- Future Proof – Modular – same data can be used by many different projects and technologies
- Platform cost – extremely attractive cost structure
Right Data Challenges: Most Data is Lost, So it Can’t Be Analyzed Later

Data Exhaust

Only a small portion of data in enterprises today is saved in data warehouses
Right Data: Save Raw Data Now to Analyze Later

- You don’t know now what data will be needed later
- Save as much data as possible now to analyze later
Right Data: Save Raw Data Now to Analyze Later

• Don’t know **now** what data will be needed later

• Save as much data as possible now to analyze **later**

• Save **raw** data, so it can be treated correctly for each use case
Right Data Challenges: Data Silos and Data Hoarding

• Departments hoard and protect their data and do not share it with the rest of the enterprise

• Frictionless ingestion does not depend on data owners
Right Interface: Key to Broad Adoption

• Data marketplace for data self-service

• Providing data at the right level of expertise
<table>
<thead>
<tr>
<th>Clean, trusted, prepared data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
</tr>
</tbody>
</table>

Data Scientists | Business Analyst
Providing Data at the Right Level of Expertise

Clean, trusted, prepared data

Raw data

Data Scientists

Business Analyst
Providing Data at the Right Level of Expertise

Clean, trusted, prepared data

Raw data

Data Scientists

Business Analyst
Roadmap to Data Lake Success

Organize the lake

Set up for Self-Service

Open the lake to the users
Organize the Data Lake into Zones
Multi-modal IT – Different Governance Levels for Different Zones

- **Raw or Landing**
  - Minimal governance
  - Make sure there is no sensitive data
  - *Data Engineers*

- **Sensitive**
  - Heavy governance
  - Restricted access
  - *Data Stewards*

- **Gold or Curated**
  - Minimal governance
  - Make sure there is no sensitive data
  - *Data Scientists*
  - *Business Analysts*

- **Work**
  - Heavy governance
  - Trusted, curated data
  - Lineage, data quality
Business Analyst Self-Service Workflow

Find and Understand → Provision → Prep → Analyze

Set up for Self-Service
Finding, understanding and governing data in a data lake is like shopping at a flea market.

“We have 100 million fields of data – how can anyone find or trust anything?” – AT&T Executive
I need data to use with self-service tools but I can’t explore everything manually to find and understand it.

I can’t govern and trust the data (metadata, data quality, PII, data lineage).

I can’t inventory all the data manually and keep up with data provisioning.
Imagine shopping on Amazon.com – an Online Marketplace
Waterline Data is like Amazon for Data in Hadoop – an Enterprise Data Marketplace
Finding and Understanding Data

• Crowdsources metadata and automate creation of a catalog
• Institutionalize tribal data knowledge
• Automate discovery to cover all data sets
• Establish trust
  • Curated annotated data sets
  • Lineage
  • Data quality
  • Governance
Accessing and Provisioning Data

You cannot give all access to all users
You must protect PII data and sensitive business information

Top down approach

• Find and de-identify all sensitive data
• Provide access to every user for every dataset as needed

Agile/Self-Service Approach

• Create a metadata-only catalog
• When users request access, data is de-identified and provisioned
Provide a Data Marketplace Interface to Find, Understand and Provision Data
Data Prep

Prepare data for analytics
- **Clean data**
  - Remove or fix bad data, fill in missing values, convert to common units of measure
- **Shape data**
  - Combine (join, concatenate)
  - Resolve entities (e.g., create a single customer record from multiple records or sources)
  - Transform (aggregate, filter, bucketize, convert codes to names, etc.)
- **Blend data** - harmonize data from multiple sources to a common schema/model

Tooling
- Many great dedicated data wrangling tools on the horizon
- Some capabilities in BI/data visualization tools
- SQL and scripting languages for the more technical analysts
Data Analysis

- Many wonderful self-service BI and data visualization tools
- Mature space with many established and innovative vendors
"Without data discovery accelerators (like Waterline Data), it may be less practical to open up Hadoop-based data hubs to business users to explore and use on their own."  
Boris Evelson, Boost your Business Insight by Converging Big Data and BI
A Successful Data Lake

Right Platform + Right Data + Right Interface
Quick Overview of the MapR Converged Data Platform

• Broad analytics capabilities

• Interoperability Standards-based APIs + POSIX NFS

• Business continuity HA with no complex configurations, incremental mirroring, consistent snapshots

• Cost effectiveness Higher performance, simplified stack, transparent compression, distributed master (NameNode) data

• Multi-tenancy capabilities Volumes, data/job placement control, granular security