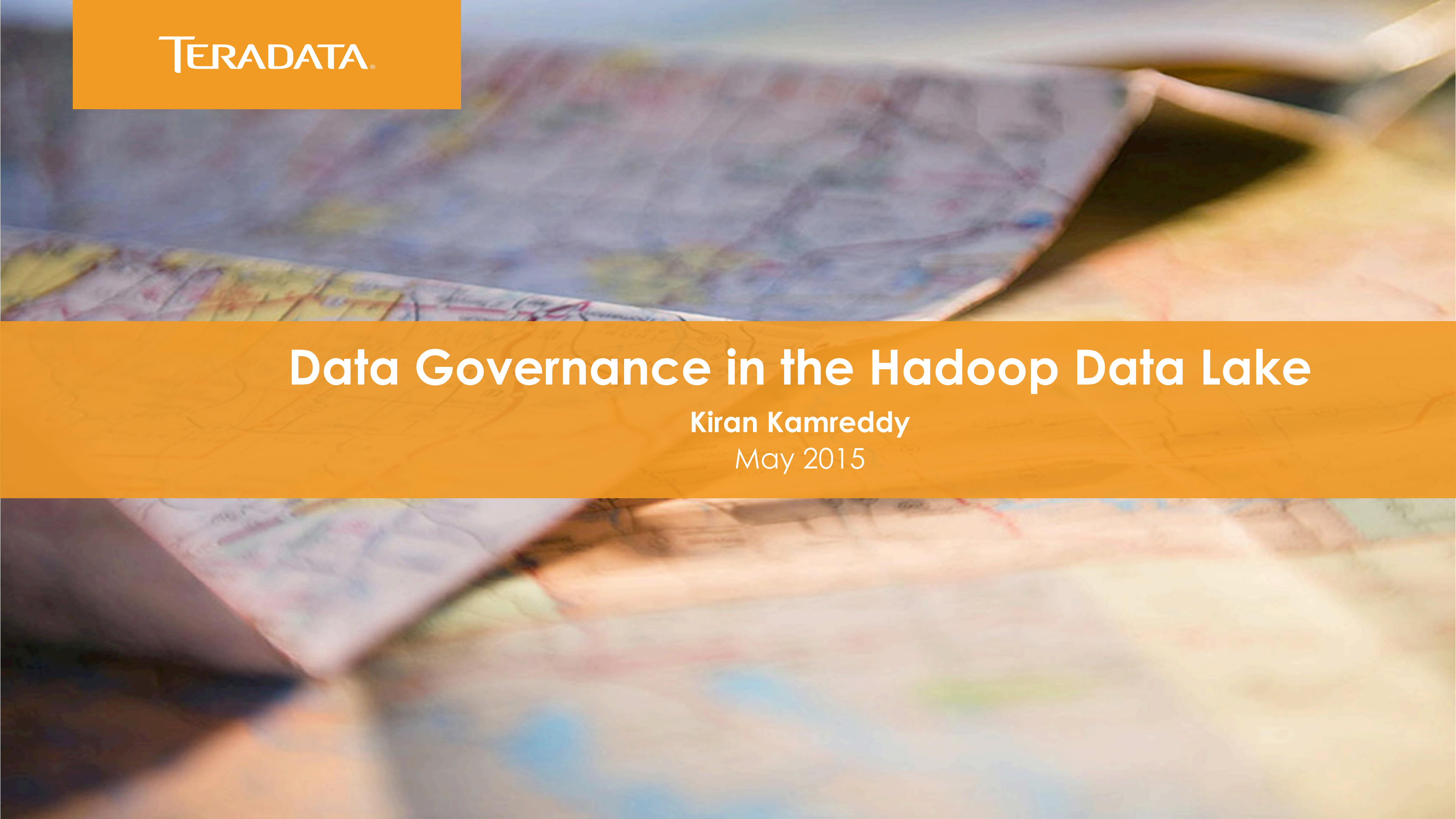


TERADATA

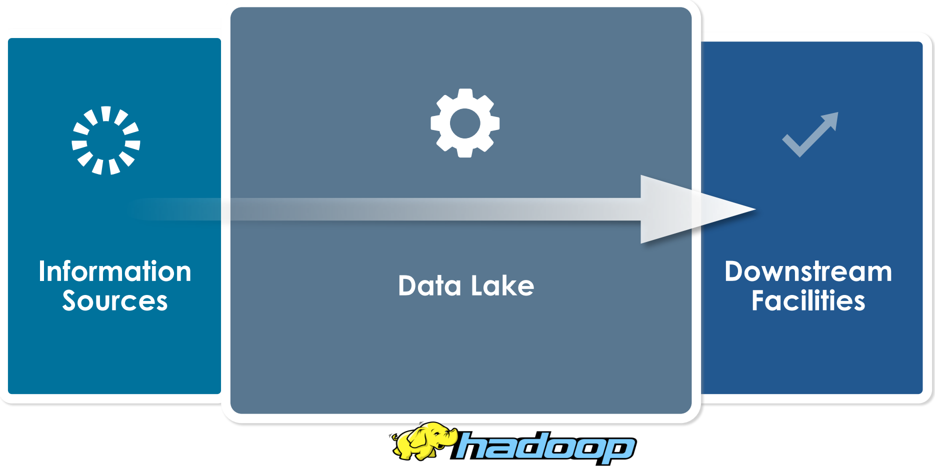
# Data Governance in the Hadoop Data Lake

Kiran Kamreddy

May 2015

# One Data Lake: Many Definitions

A **centralized** repository of **raw** data into which **many** data-producing streams flow and from which downstream facilities may draw.



**Data Variety** is the driving factor in building a Data Lake

# Data Lake Maturity & Risks

**Data Lake initiatives usually start small**

- More Data Sources
- More Applications
- More Business Units
- More Users

**…. grow into more complex environments**

**Without proper governance mechanisms**

**Data lakes risk turning data swamps**

# What Is Data Governance ? And What can it do for my Data Lake

**Fundamental capabilities for organizing, managing and understanding data**

- Where did my data come from ? How is it being transformed ?
- Track usage, resolve anomalies, visualize, optimize and clarify data lineage
- Search and access data ( not only browse )
- Assess data quality and fitness for purpose

**Specialized capabilities to meet regulatory/compliance requirements**

- Govern who can/cannot access the data and who cannot
- Data life cycle management, archiving and retention policies
- Auditing, compliance

**Data Governance first approach to prevent turning to Data swamps**
**Retrofitting data governance is not feasible**

# Governance and Productivity

**Governance should supports day-to-day use of data**

- Data workers need a strong understanding
- Roles for data stewards, data owners, data analysts/scientists need to be assigned

**Operational Metadata is critical to understanding**

- Where did it come from?
- What is the environment? – landing zone, OS, Line of Business
- What processes touched my data?did you lose any data? – Checksums etc.
- When did the data get ingested, transformed?
- Did it get exported, when, where how will it be used (organizational)?

**Provision consistent ingest methods that track operational metadata**

# What is Regulatory Compliance ?

- **Compliance and Regulatory**
  - Capture, store and move data
  - Sarbanes-Oxley, HIPAA, Basel II
- **Security**
  - Authorization, Authentication
  - Handling sensitive data
- **Auditing**
  - Recoding every attempt to access
- **Archive & Retention**
  - Data life cycle policies

## 3 Market Approaches

- Apache Hadoop has built-in support for these capabilities
- Hadoop distribution vendors have all made improvements in each of these areas
- A variety of vendors provide specialized capabilities in each area that go beyond what a Hadoop distribution provides

# Data Governance Challenges on Hadoop

## Hadoop is different to DW

- **Scale**: High volumes of data, multiple user access
- **Variety**: Schema-on-read, multiple formats of data
- Multiple storage layers (HDFS, Hive, HBase)
- Many processing engines (MR, Hive, Pig, Impala, Drill…)
- Many workflow engines/schedules (Cron, Oozie, Falcon…)
- Holistic view of data with required context is difficult

## Hadoop needs less stringent, more flexible mechanisms

Balance agility and self service with processes, rules, regulations

Maintain Governance without losing Hadoop's power

# Teradata's Approach for Data Governance in Hadoop

**Teradata Loom® – Integrated Data Management for Hadoop**

Metadata management, Lineage, Data Wrangling

Automatic data cataloging, data profiling and statistics generation

**Teradata Rainstor – Data Archiving**

Structured data archiving in Hadoop with robust security

Compliance and auditing

**ThinkBig – Hadoop professional services**

Hadoop Data Lake – packaged service/product offering to build and deploy high-quality, governed data lakes

# Teradata Loom®

## Find and Understand Your Data

- ActiveScan
  - Data cataloging
  - Event triggers
  - Job detection and lineage creation
  - Data profiling (statistics)
- Workbench and Metadata Registry
  - Data exploration and discovery
  - Technical and business metadata
  - Data sampling and previews
  - Lineage relationships
  - Search over metadata
  - REST API – easily integrate third-party apps

## Prepare Your Data

- Data Wrangling
  - Self-service, interactive data wrangling for Hadoop
  - Metadata tracked
- HiveQL
  - Joins, unions, aggregations, UDFs
  - Metadata tracked in Loom

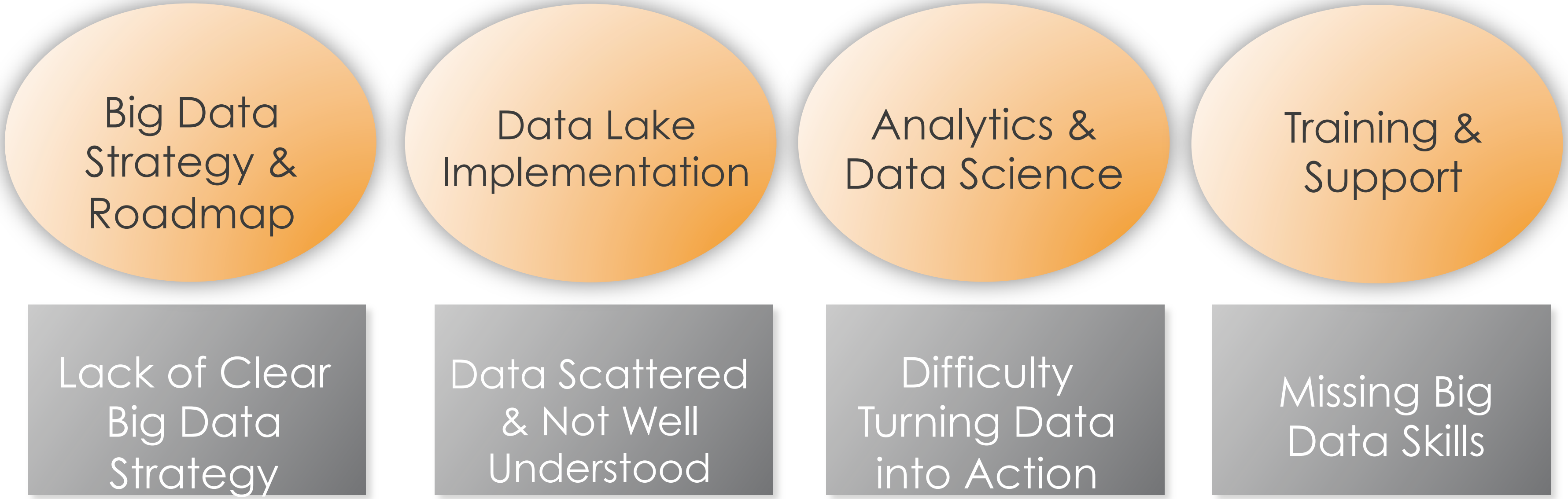
# Teradata RainStor

- Retain data online that is **queryable** for an indefinite period
- Retire data that are no longer required with **auto-expiration** policies
- **Comply** with strict government rules and regulations
- Retain the **metadata** as it was originally captured
- Store tamper-proof, **immutable** (unchangeable) data
- Maintain availability to data as RDBMS versions change or expire
- Compression, MPP SQL query engine, Encryption, Auditing

# Think Big Data Lake Starter

- Enables a rapid build for an initial Data Lake
- Data Lake Build - Provide recommendations and assistance in "standing up" a 8-16 node data lake on premises or in the cloud
  - Implement and document 2-3 Ingest Pipelines
  - Robust infrastructure to support fast onboarding of new pipelines and use cases
  - Implement an end-to-end Security Plan
    - Perimeter, authentication, authorization and protection
  - Integrated data cataloging and lineage through Loom
  - Implement archiving, if required, through RainStor

# Big Data Services from Think Big



Big Data Strategy & Roadmap

Data Lake Implementation

Analytics & Data Science

Training & Support

Lack of Clear Big Data Strategy

Data Scattered & Not Well Understood

Difficulty Turning Data into Action

Missing Big Data Skills

Focused exclusively on tying Hadoop and big data solutions to measurable business value

# Data Governance for Hadoop
# Bank Holding Company

## Situation

Large scale data lake planned with many heterogeneous sources and many individual analyst users.

## Problem

Lack of centralized metadata repository makes data governance impossible. Enterprise must have transparency into data in the cluster and capability to define extensible metadata.

## Solution

Hadoop provides data lake infrastructure. Loom provides centralized metadata management, with an automation framework.

**Impact**

- **Co-location of data provides more efficient workflow for analysts**
- **Hadoop provides scalability at a lower cost than traditional systems**
- **Develop new insights to drive business value**

# Summary

- **Data governance is critical to building a successful data lake**
  - Fundamental governance capabilities make data workers more productive
  - Solutions for meeting regulatory requirements are also needed
- **Teradata Loom provides required data cataloging and lineage capabilities to make hadoop users more productive**
- **RainStor provides advanced archiving solution**
- **ThinkBig Data Lake provides the complete package**

**Stop by Our Booth for a Demo**

TERADATA®