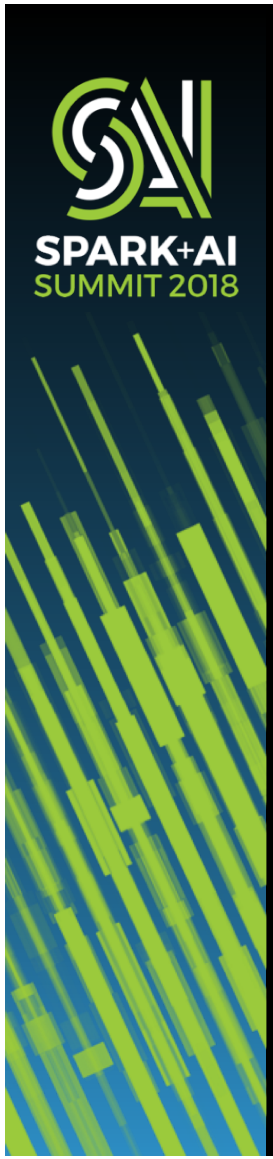
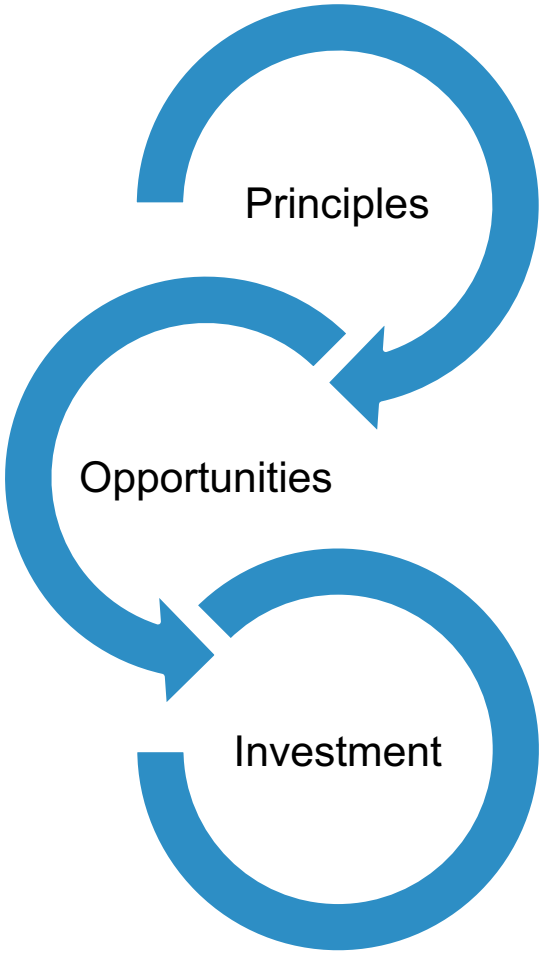


Evolution of eBay's Enterprise Data Ecosystem with Apache Spark

Kim Curtis, Brian Knauss, eBay

#EntSAIS13

Evolution



Principles

- Expand Capabilities
- Increase Flexibility
- Optimize Cost/Performance

Principles

DO MORE

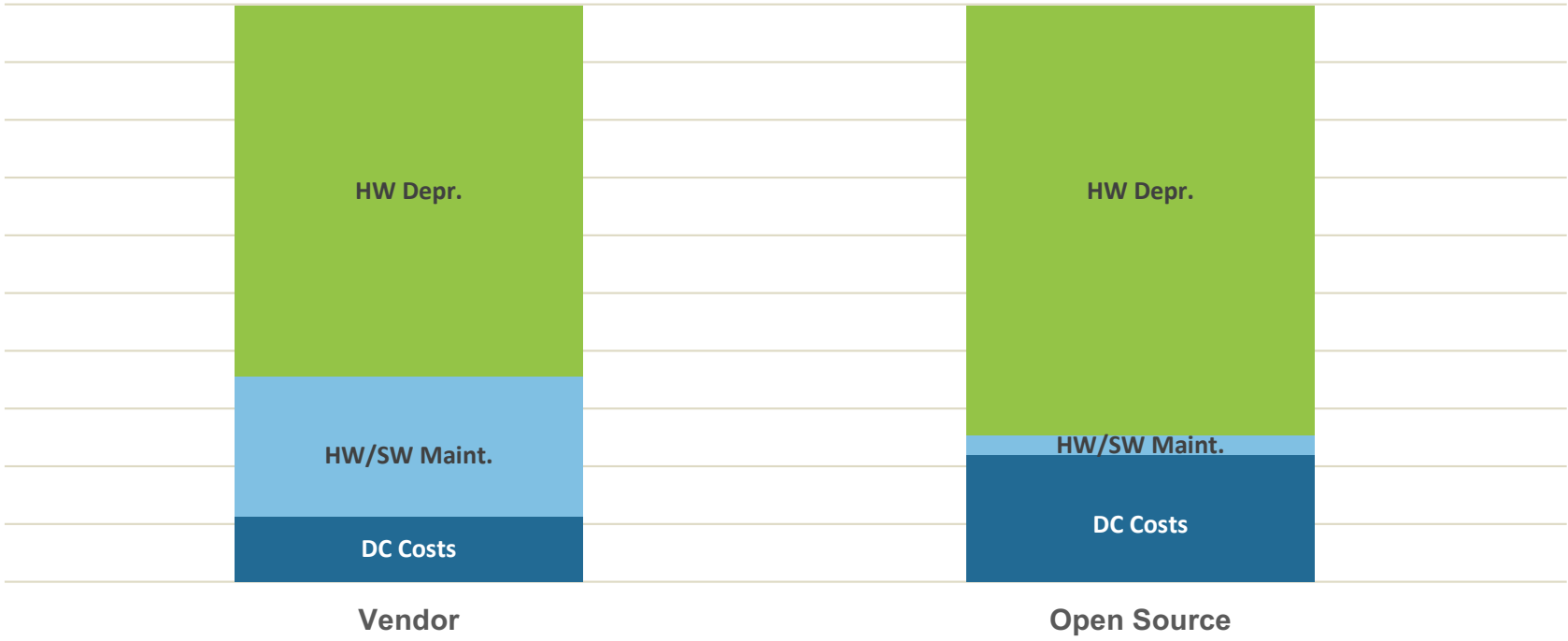
WITH FEWER BARRIERS

CHEAPER/FASTER

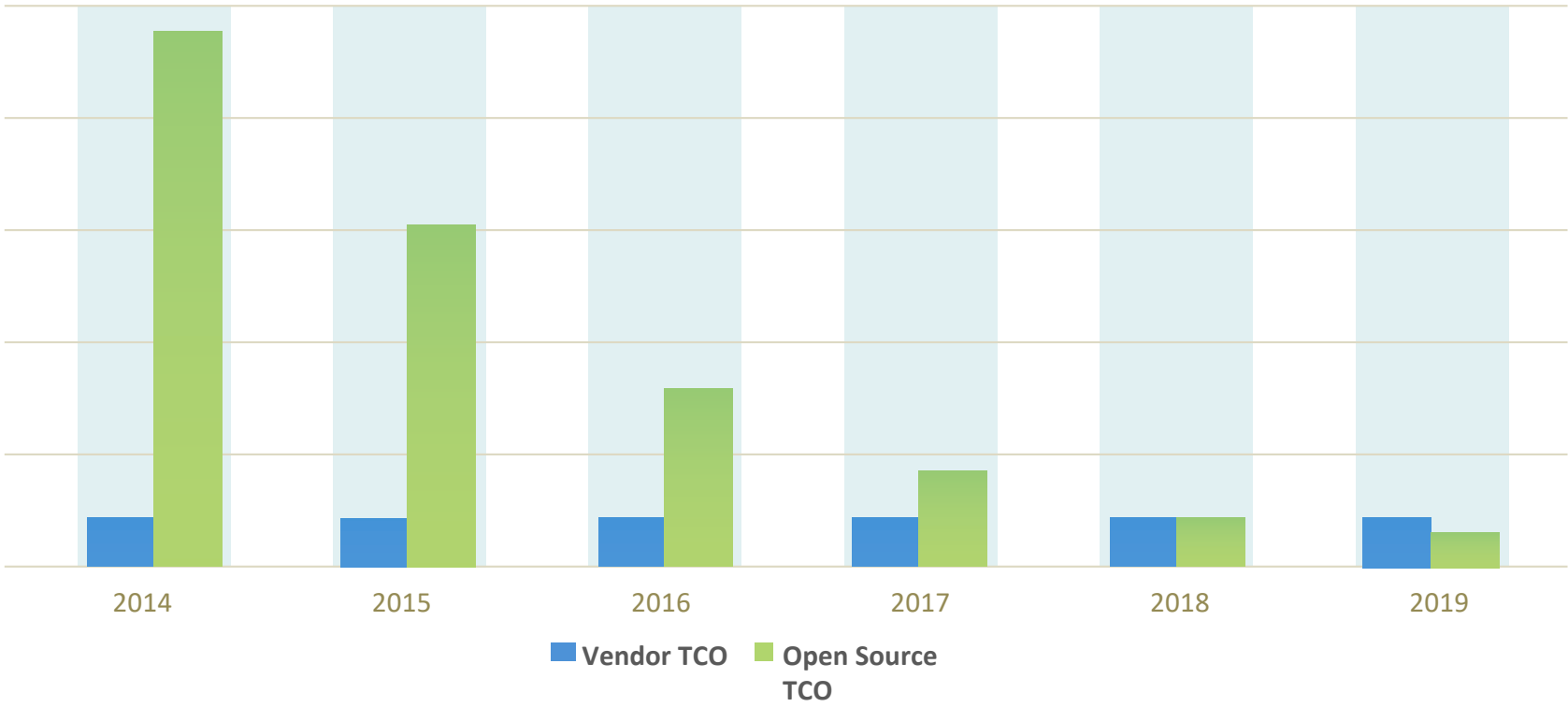
Spark in the EDW

- Increase Flexibility
  - Investment and Engineering
- Expand Capabilities
  - Pre-Load Transformation
- Optimize Cost/Performance
  - Let's see…

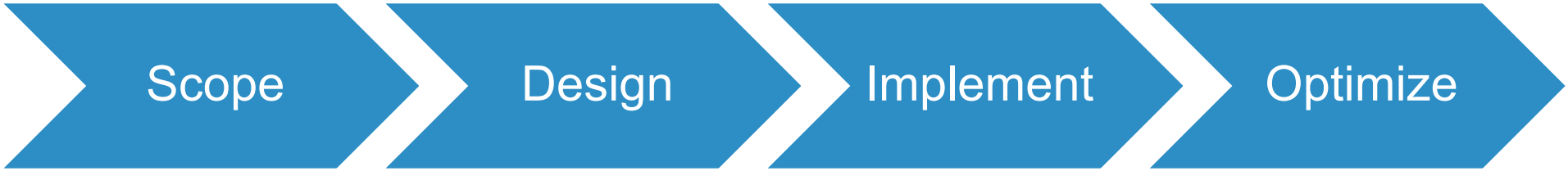
TCO Comparison



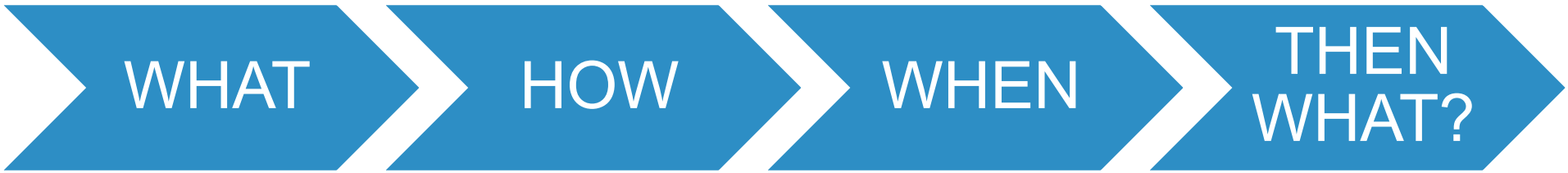
Opportunity Assessment



Investment



DO IT



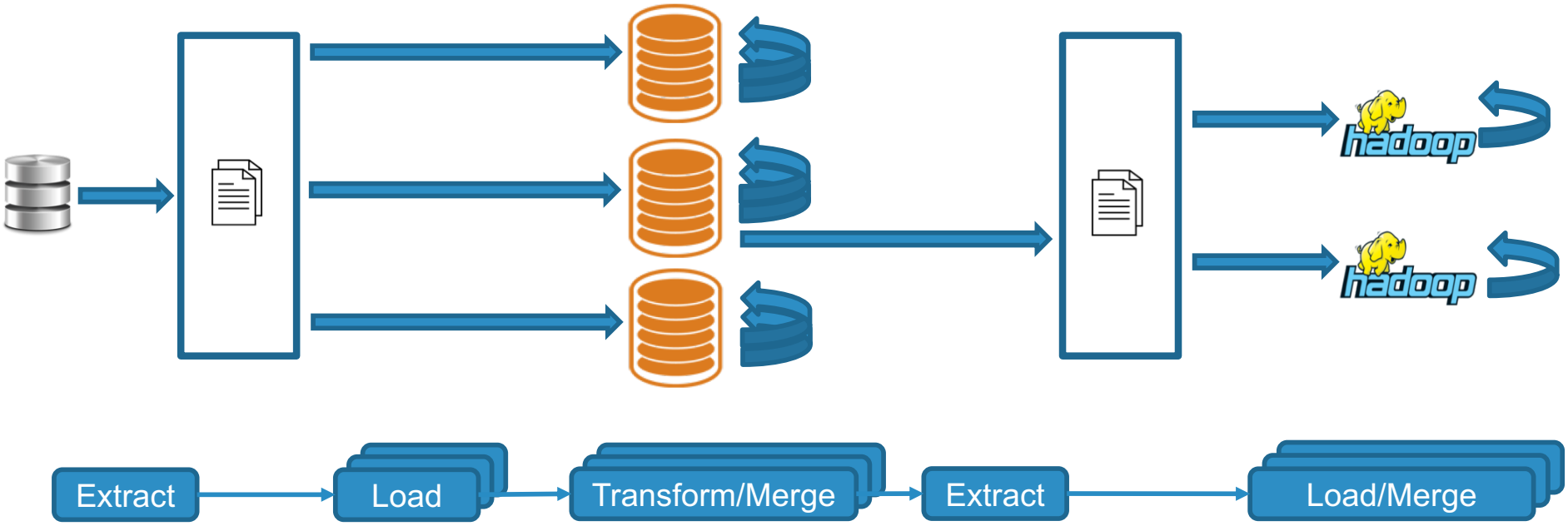
Scope

- Engage with Customers
  - Isolated our impact
- Define Boundaries
  - Relational Batch Processing
- Set Intermediate Targets
  - Offset 2017 growth

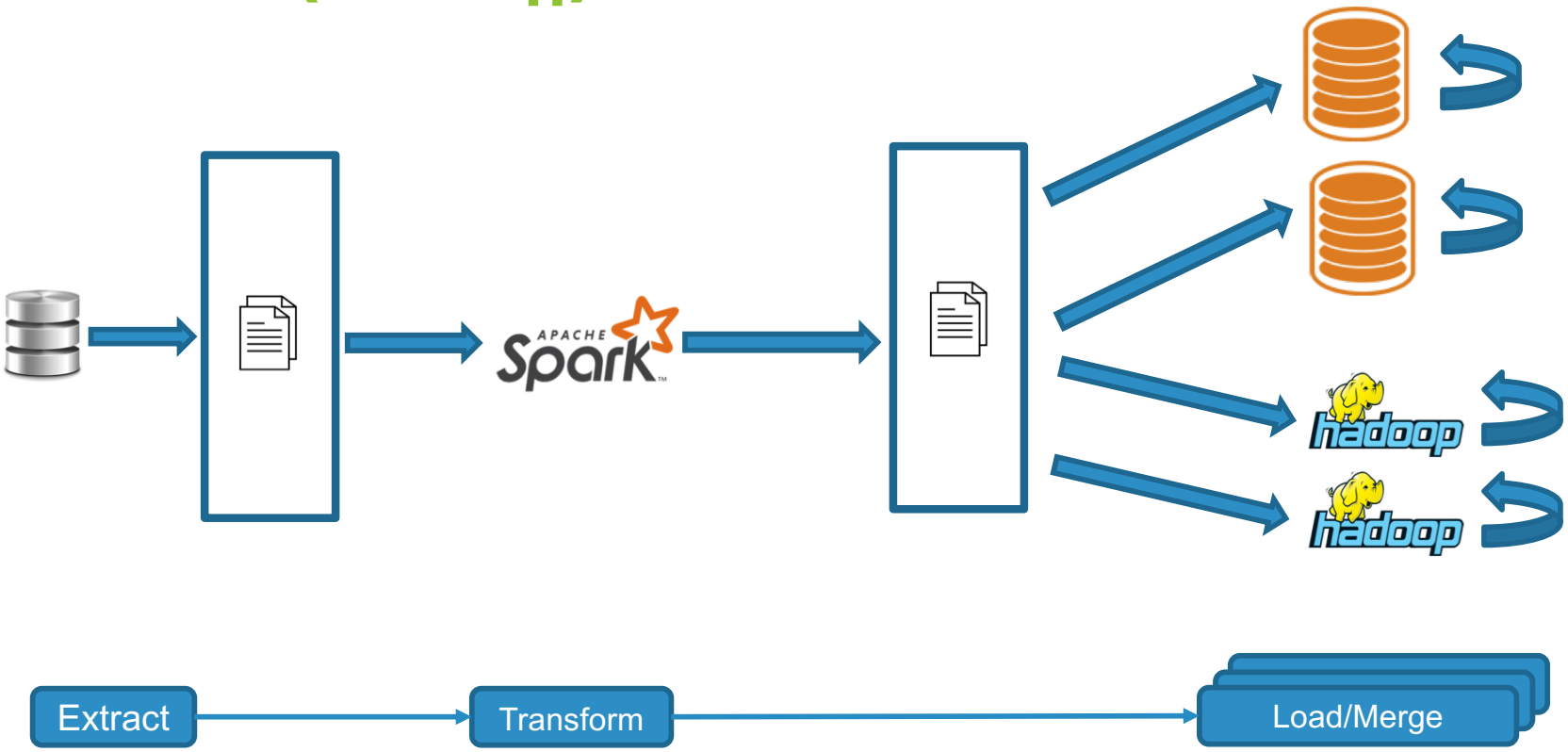
Design

- Extensible Framework
  - (EL_nT_n to ETL_n)
- Optimize Hardware
  - Processing and Storage Nodes
- Optimize Software
  - Automated Spark SQL Tuning

Before (EL_nT_n)



After (ETL_n)



Implement

- Production HDFS Data Environment
  - Tight Alignment with Platform
- Prioritized Effort
  - Minimize effort to hit goals
- Distributed Engineering (internal Open Source)
  - Built and tested framework additions

Optimize

- Scaling/DR(DA)
  - Multi-platform architecture
- Cost/Performance Optimization
  - HW/SW tuning
- Feature Expansion
  - ML on platform

Challenges

- Scale of Data and Workload
  - Data Validation Between Targets
- Migration Automation
  - Migration of Non-Standard Processes
- Enterprise Readiness of Open Source
  - Job-Level Workload Tracking/Management

The ~~End~~ Beginning…