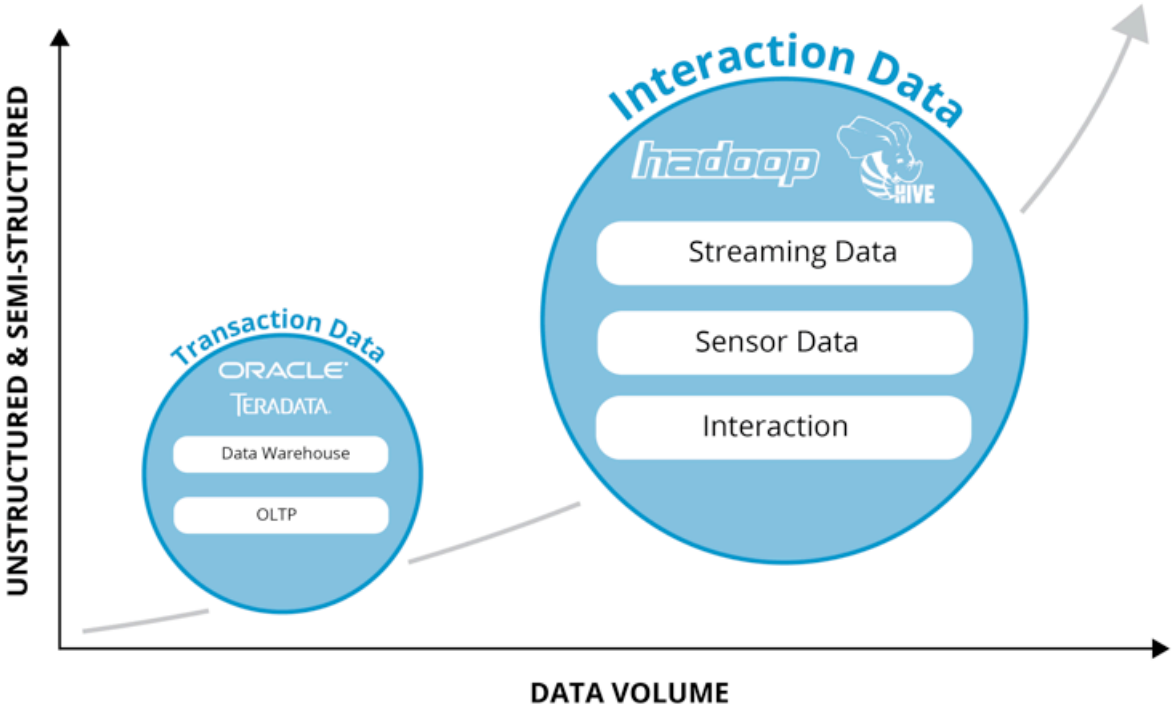
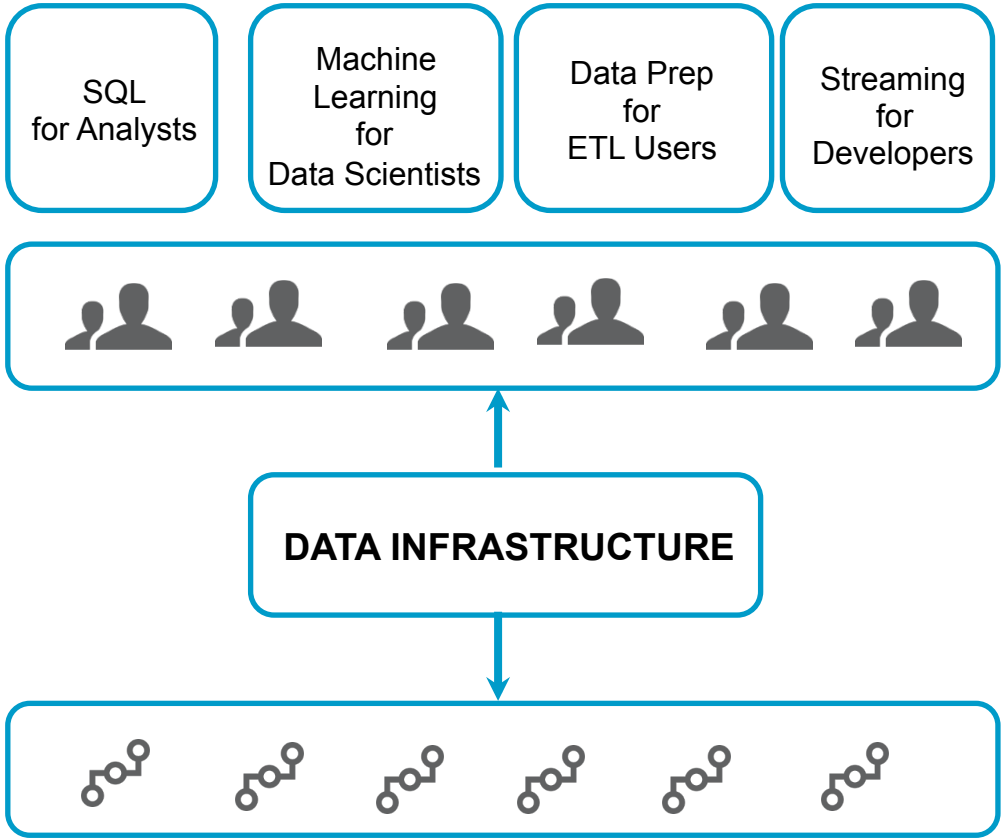
Data Lifecycles at Massive Scale

Pasadena, January 2016

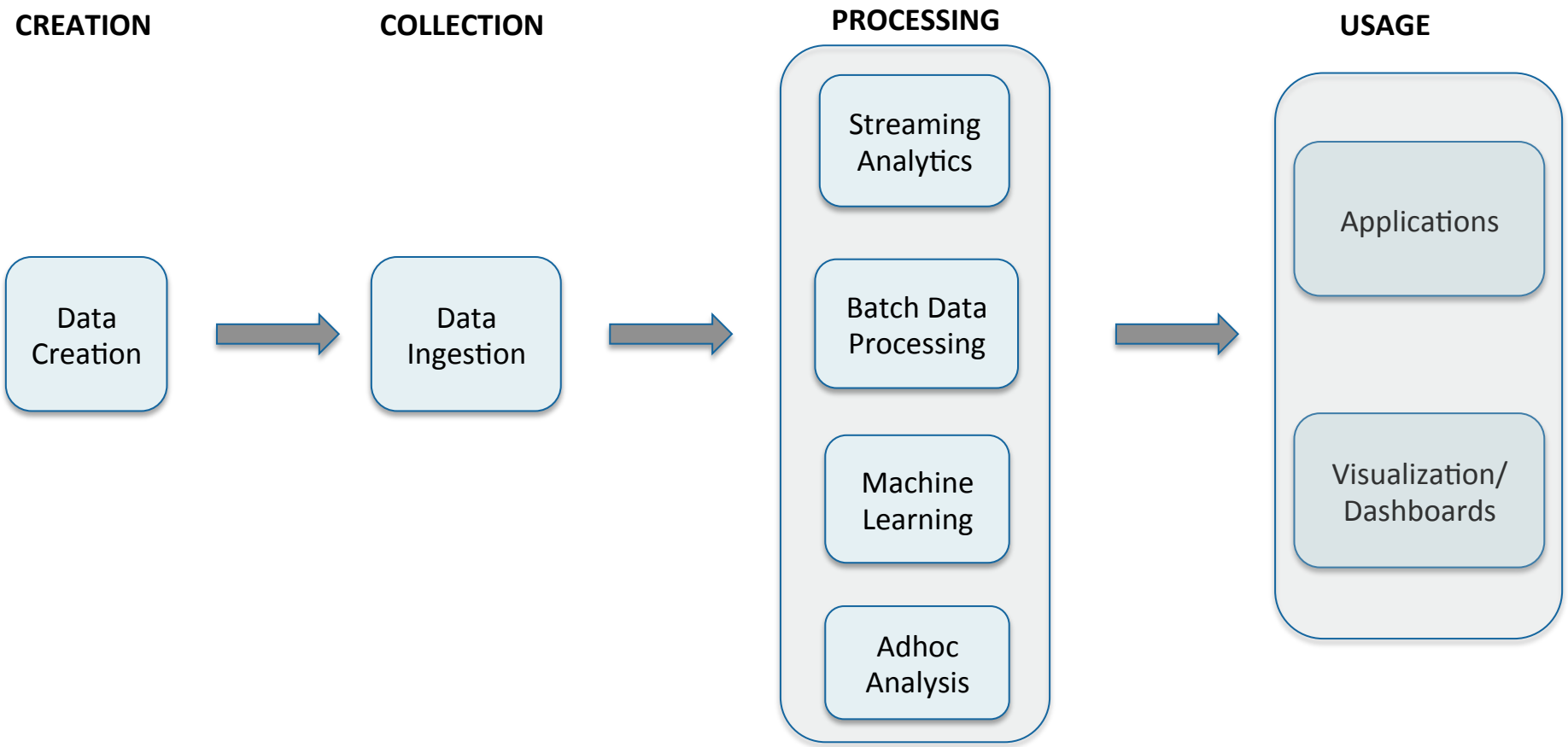
- Co-founder and CEO of Qubole
  - Cloud Based Big Data as a Service
  - Processes 250PB+ data every month
- Lead Data Infrastructure at Facebook
  - Made Big Data Self-service in Facebook
  - Nearly an Exabyte of data
- Co-creator of Apache Hive
  - Democratized Big Data through SQL

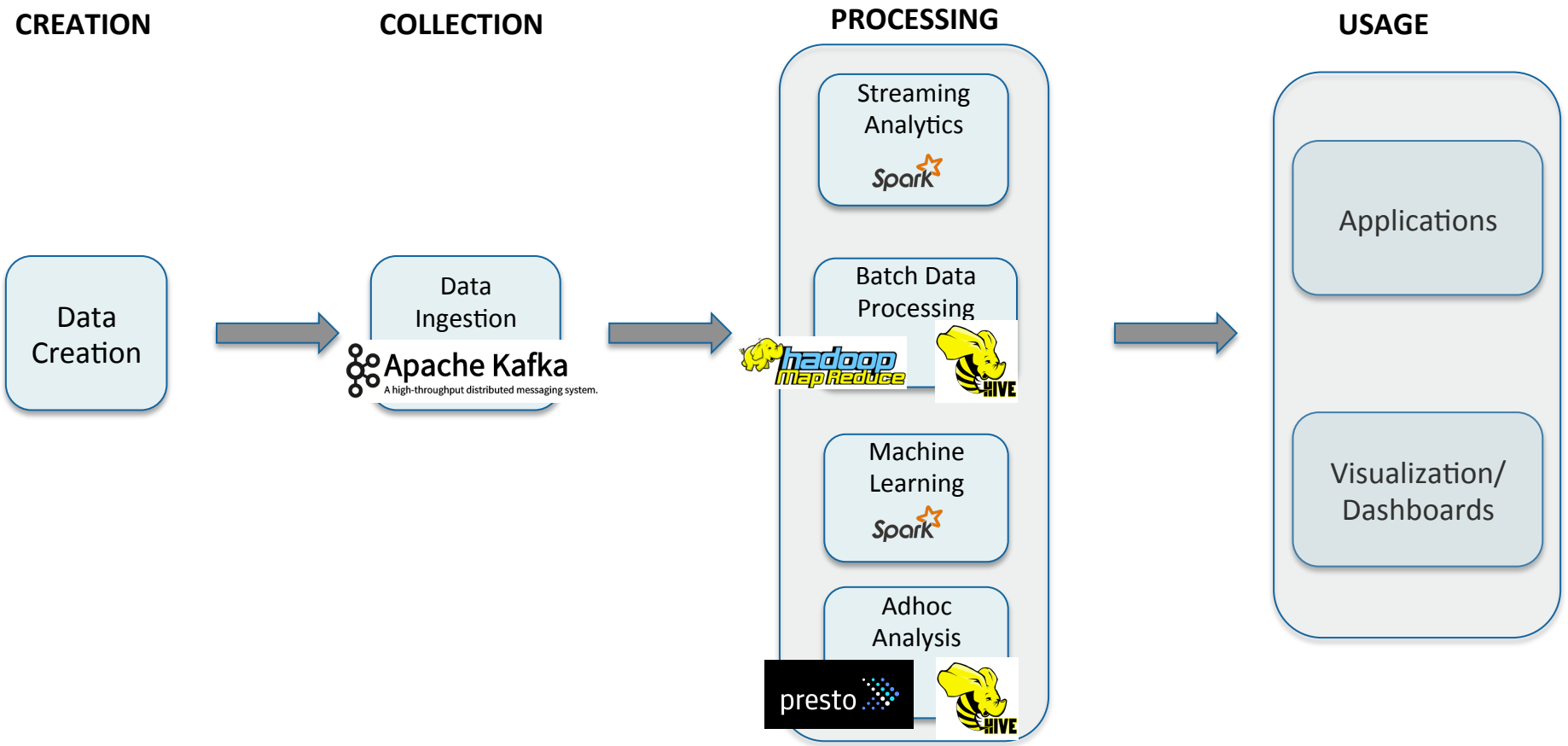

Interaction Data
hadoop
HIVE
Streaming Data
Sensor Data
Interaction
Transaction Data
ORACLE
TERADATA
Data Warehouse
OLTP
UNSTRUCTURED & SEMI-STRUCTURED
DATA VOLUME



Multi-Persona support for Multiple Use-cases

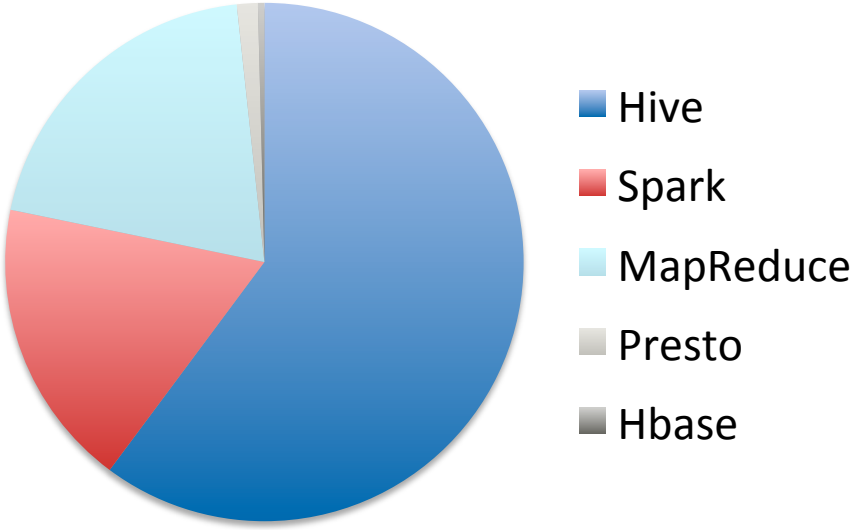
Scalability on Commodity Hardware


CREATION
COLLECTION
PROCESSING
USAGE
Data Creation
Data Ingestion
Streaming Analytics
Batch Data Processing
Machine Learning
Adhoc Analysis
Applications
Visualization/ Dashboards


CREATION
COLLECTION
PROCESSING
USAGE
Data Creation
Data Ingestion
Apache Kafka
A high-throughput distributed messaging system.
Streaming Analytics
Spark
Batch Data Processing
hadoop MapReduce
HIVE
Machine Learning
Spark
Adhoc Analysis
presto
HIVE
Applications
Visualization/ Dashboards

250PB+ Data Processed Every Month

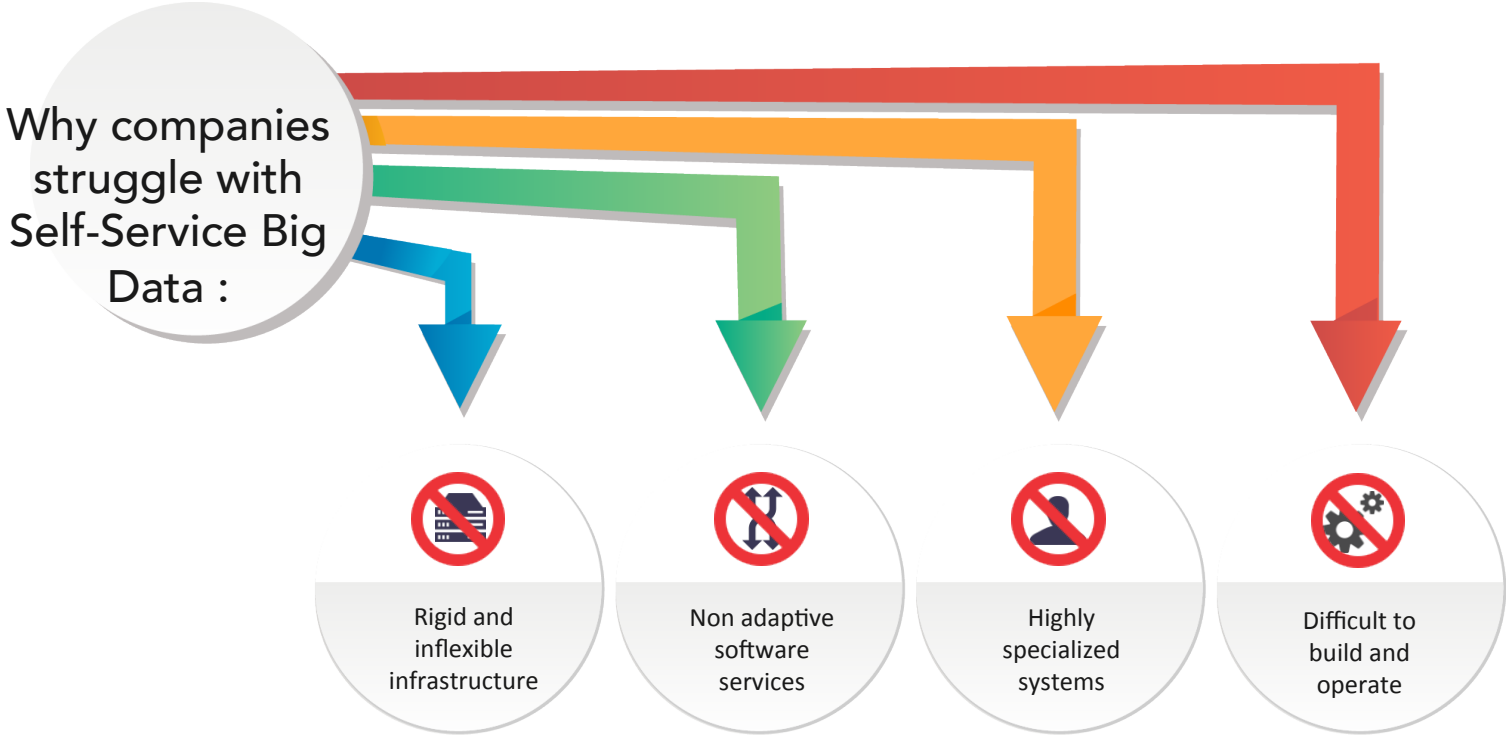
Engine Usage on Qubole



A Modern Data Platform needs Multiple Engines

- Hive for Complex SQL
- Spark for Data Science and Streaming
- Presto for Interactive Simple SQL
- Map Reduce for Batch ETL

Impediments for an Aspiring Data Driven Enterprise



- **6-18 month** implementation time
- **Only 27% of Big Data initiatives** are classified as "Successful" in 2014
- **Only 13% of organizations** achieve full-scale production
- **57% of organizations** cite skills gap as a major inhibitor

The Cloud Provides:

On-demand
Infrastructure

Highly Scalable
Object Stores

Self Service
Infrastructure

The Right Storage for Storing Big Data

- **Elastic Scalability**: petabyte scale without capacity constraints
- **High Concurrency**: throughput scales linearly
- **High Availability**: 99.99% availability
- **High Durability**: easily more durable than HDFS 3x replicas
- **Enterprise Grade** at a **fraction** of the cost

The Right Compute Paradigm to Fit Usage

- **On-Demand**: provision entire clusters in less than two minutes with no lead time for sourcing
- **Elastic**: cluster size should match workload; run with thousands of nodes when you need it, de-provision all nodes when idle
- **Flexible**: change compute infrastructure to fit workloads
- **Cost Efficiency**: Multiple SLA options to fit the right budget to the workloads

Big Data Meets Cloud – Power of Self Service SaaS

Qubole

On-demand elastic clusters

Integrated big data software

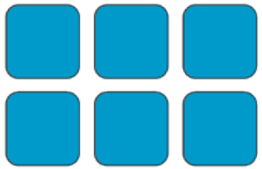
Benefits:

- Agile platform – 10 to 1000 nodes in minutes
- Flexible infrastructure – different types of nodes of different work loads
- Zero Operations
- Lower TCO

DIY

(Cloudera, Hortonworks etc.)

Static clusters



Big data component confusion

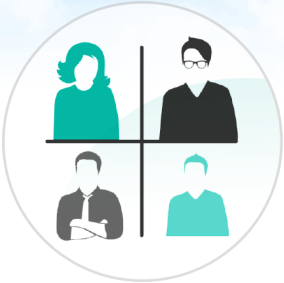
Cons:

- Lot of planning to get clusters up and running
- Inflexible and static infrastructure
- Need Hadoop operations experts
- Higher TCO

Security is a no. 1 citizen: Cloud Built from Outside–In

- Multiple Encryption options
- Industry-standard authentication for every REST API request
- Virtual Private Cloud
- Auto-logging for auditability
- Industry Compliance

Successful Big Data Adoption at Scale with a Unified Big Data Platform Built for the Cloud

Multi-User and SaaS architecture for Best Operational Efficiency

Enterprise Grade Security, Governance and Reliability

Auto-scaling and portability across Clouds

Thank You