Big Data on Tap

cask.co

Building an Enterprise Data Lake

Jonathan Gray
Founder & CEO
@jgrayla

Data Summit, New York
May 16, 2017
The Promise of Big Data

Big Data
- Digitization
- Internet of Things
- Social and Mobile

Business Needed
- Age of the Customer
- Data-driven Culture
- Digital Disruption

Skills
- Data Science
- Machine Learning
- Predictive Analytics

Powerful IT
- Commodity
- Cloud
- Open Source

Customer Service & Engagement
- Market Differentiation
- Business Optimization
- Financial performance & risk control

The Promise of Big Data
But Getting Value from Big Data is Difficult

This Stuff is Hard
- Projects not Products
- Lack of Big Data Talent
- Heroics not Blueprints

Data Integration is Challenging
- Integration is complex & time-consuming
- Risks of creating a data swamp
- Point solutions (often) result in shadow IT

Not Enterprise-Ready
- Divergent & disaggregating stacks
- Security and access control risks
- Lack of governance capabilities

“Through 2018, 70% of Hadoop installations will fail to meet goals for cost savings and revenue generation objectives due to skills and integration challenges …”
What is a **Data Lake**

**James Dixon (Pentaho)**

Data Lake

Data streams in from sources to fill the lake, and various users of the lake can come to examine, dive in, or sample. *(Hadoop World NYC 2010)*

**Gartner**

Data Lake

Enterprise-wide data management platforms for analyzing disparate sources of data in its native format.

**Cloudera**

Data Hub

A centralized, unified data source that can quickly provide diverse business users with the information they need to do their jobs.

**Hortonworks**

Data Lake

Collect everything, dive in anywhere, give flexible access. Maximum scale and insight with the lowest possible friction and cost.

**Cask**

Data Lake

Any user has self-service access to any data, with the required security, policy, and governance capabilities built-in.
Vision of the Enterprise Data Lake

- Business Analyst
- Data Scientist
- App Developer
- Data Engineer
- Data Tracker

Data Lake

- Fast ingestion, storage & compute
- High performance analysis
- Data governance & federation

Categories of data:
- Sensor data
- Historian data
- Machine data
- CRM, ERP, etc.
- Logs
- Geo-location data
- Social network data
The Big Data Technology Stack enables but *does not deliver a Data Lake*
Data Lake Architectures

**Data Pond**
Raw Data copied from existing internal data stores and pulled from external data sources [*pre-production*]

**Data Lake**
Raw + Defined Data regularly pushed from other systems into a centralized cluster [*production*]

**Data Reservoir**
Raw + Defined Data which is governed and audited to ensure compliance and security [*multi-tenant*]
Data Lake Challenges

**Data Pond**
- **Manual processes** requiring hand-coding and reliance on command-line tools
- **Hard to find data** and it’s lineage for data discovery and exploration
- **Coupling of ingestion and processing** drives architecture decisions

**Data Lake**
- **Operationalizing** processes for production and to maintain SLAs
- **Ensuring data is in canonical forms** with a shared schema usable by others
- **Coding or filing tickets** often required to perform new ingestion and processing tasks

**Data Reservoir**
- **Multiple architectures** and technologies used by different teams on different clusters
- **Guaranteeing compliance** in a system that is designed for schema-on-read and raw data
- **Sharing infrastructure** in a multi-tenant environment without low-level QoS support
Enter Cask

**Founded in 2011**
By early Hadoop engineers from Facebook and Yahoo!

**Strategic Investors**
AT&T, Cloudera and Ericsson

**Key Customers & Partners**
AT&T, Cloudera, Ericsson, Hortonworks, IBM, Lotame, MapR, Microsoft, Salesforce, Tableau...

**Unique Value Proposition**
First Unified Integration Platform for rapid time-to-value from Big Data

**Mature Platform: CDAP 4**
Featuring Cask Market, the “big data app store”

**Why “Cask” ?**
A Container Architecture that puts Big Data on Tap

**Raised $37+ Million**
Andreessen Horowitz, Safeguard, Battery Venture and Ignition Partners
CDAP: END-TO-END UNIFIED INTEGRATION PLATFORM

Data Integration
- Modern Data Ingest, Data Prep, Pipelines and Metadata Management

App Dev
- Distributed Application Server for Large Scale Data-Driven Apps, Analytics, and Services

Security & Governance
- End-to-end security and automated tracking of lineage and access

Self-Service
- Rich, interactive UIs with drag-and-drop workflows for broad, non-developer user access

Portable & Production
- Move rapidly into production with automatic support for new environments and technologies

Unified data management and application development for data lakes and data apps
The Evolution of the Cask Platform

Convergence of Big Data Apps and Data Integration

CDAP v2

Big Data App Server
- Abstractions & integrations
- Metrics & logs
- Debugging environment

“WebLogic for Hadoop”

CDAP v3

Big Data Apps + Data Integration
- Data ingest
- Data pipelines
- Workflows and metadata

“WebLogic Meets Informatica”

CDAP v4

Unified Integration for Big Data
- Security & governance
- Self-service environment
- Enterprise integrations

“Unified Big Data Integration”
Increased Focus on Usability and Business Outcomes

- **Solutions**
  - Completed Solutions with Interactive Wizards and UIs
  - Apps, Plugins, Drivers, Pipelines both Public and Private

- **Product**
  - Self-Service User Interfaces for Non-Developers
  - No Code, End to End Common Big Data Usage Patterns

- **Platform**
  - Extensible, Portable, Production Runtime Environment
  - Built-in Enterprise Security and Governance Capabilities
Cask Market

The Big Data “App Store”

Pre-built Components

Rapid Deployment

Packaged Solutions

Seamless Experience
Data Lakes on CDAP

**Data Pond**

CDAP and Cask Market with templates and plugins enables production workflows in minutes

Never lose data by ensuring all ingested data is tracked with metadata and provenance

Separation of ingestion and processing to support any type, format and rate

**Data Lake**

Operationalize workflows using scheduling and SLA monitoring with time / partition awareness

Using common transformations and a shared system for defining and exposing schema

Reference architecture ensures a common platform across teams, orgs, ops and security

**Data Reservoir**

Multi-tenant namespacing provides data and app isolation, tying together infrastructure

Ensure compliance by requiring the use of specific transformations and validation

Self-service access discovery, ingest, prep, processing and exploration of data
Questions?

Come see us at Booth #7 here at Data Summit!

For more information, go to cask.co or contact us at info@cask.co

Tweet us @jgrayla @CaskData