Transforming Cloudera analytics agility in a fast-moving world

Brent Compton, Sr Director Tech Marketing, Red Hat
Idan Levi, formerly at Gol, now in Office of CTO, Red Hat

May 2019
The Story of Unlocking

Data
Compute
Innovation
It’s All About Time to Market

- 4–5 Months

- 8–14 Minutes
The Data Factory Vision

Example: Machine learning pipeline
Cloudoop
Native Hadoop vs. Cloudoop

- Compute Service
- Object storage Service (S3)
Cloud Approach – Unlocking

• Elasticity
• Compute-data decoupling
• Scalability
• Dynamic
• Multi tenancy
• SLA
Cloud Approach – AWS experience on-prem

- **S3-compatible object store**
- **Persistent cluster** – interactive queries (Spark-SQL | Impala)
- **Transient cluster** – batch jobs (X hours nightly) – add/remove nodes
- **Workload specific clusters** (different sizes, different versions)
Cloud Approach
Switch Off Clusters

S3 → S3

S3 → S3

S3 → S3
Cloud Approach
Logical Separation of Jobs

Hive

S3

As you wish…

Spark job

Impala

Ad-hoc
Cloud Approach
Logical Separation of Jobs

Hive

Impala

Spark Job

Ad-Hoc

As you wish…
High Level Architecture

- Big Data Platforms
  - Cloudera

- High Level API
  - PaaS / IaaS services

- Cloud Management
  - Software-defined everything

- Low Level API
  - Hardware discovery & control

- Hardware
  - Hyper scale
Cloudera

- bare metal servers
- HDFS

Red Hat OpenStack

Red Hat Ceph
Proof Of Concept
Fine tune – Spark Terasort Before & After

Spark Terasort 500g

Elapsed time [Sec]

- Native HDFS: 1126.3158
- Cloudoop S3 After Iteration: 1586.83
- Cloudoop S3 Before: 7894

* Lower is better
Fine tune - TPC-DS Before & After

TPC-DS 500g

Native HDFS: 3647.8571
Cloudoop S3 After: 3518.83
Cloudoop S3 Before: 4699.25

*Lower is better
Testimonials

Lower is better

Running on local SSDs
Cloudoop – Phase 1

• Versions
  • CDH 5.13
  • Spark 2.2

• APIs
  • Create cluster
  • Delete cluster
  • List clusters
  • Describe cluster

• Fine tuned cluster with full management access
The Data Factory Vision

Example: Machine learning pipeline

Data Sources

Object Storage (Ceph)

Cloudoop

GaaS

Compute

Insights
Strategic partnership

Cloudera

Red Hat

& MORE
FIND US AT RED HAT SUMMIT

- At the Storage lockers
- At the Red Hat booth
- At one of Storage dedicated sessions (red.ht/storageatsummit)
- At the Community Happy Hour (Tues 6:30, Harpoon Brewery)
- At the Hybrid Cloud Party (Wed, 7:30, “Committee” restaurant)

redhat.com/storage
@redhatstorage
redhatstorage.redhat.com
THANK YOU

linkedin.com/company/Red-Hat
youtube.com/user/RedHatVideos
facebook.com/RedHatInc
twitter.com/RedHat