



One Hadoop, Multiple Clouds

Andrei Savu | Tech Lead, Cloudera Director



About me

Tech Lead on [Cloudera Director](#)

Previously founder of [axemblr.com](#)

Contributed to Apache Whirr (PMC) & jclouds.

Twitter: <https://twitter.com/andreisavu>

LinkedIn: <https://www.linkedin.com/in/sandrei>

Cloudera Director
cloudera.com/director

Deploy and manage
enterprise-grade
Hadoop in the cloud

AWS & Google Cloud
Extensible via plugins

The screenshot displays the Cloudera Director web interface in a browser window. The page title is "cloudera director". The breadcrumb navigation shows the path: "Environment - Production AWS / Cloudera Manager - CM-1 / Cluster - CDH-1 Ad Hoc Analytics". The user is logged in as "admin".

The main section is titled "Cluster CDH-1 Ad Hoc Analytics" with a "Modify Cluster" button. Below this is a "Details" section with the following information:

Details			
Status	Ready	URL	Go to Cluster
CDH	5	Parcel Repositories	http://archive.cloudera.com/search/parcels/latest/ http://archive.cloudera.com/cdh4/parcels/latest/ + Show More

Below the details section are two tables: "Services" and "Instance groups".

Services

Other
<input checked="" type="radio"/> HDFS
<input checked="" type="radio"/> YARN
<input checked="" type="radio"/> ZooKeeper
<input checked="" type="radio"/> Spark

Instance groups

Name	Roles	Instance Template	Instance Count
compute	View Roles	View Template	15
masters	View Roles	View Template	1
workers	View Roles	View Template	10



cloudera

Journey to the Cloud

Do you use a public or private cloud?

How do you run and manage Hadoop?

cloudera



What is this talk
about?

State of the World
Architectural Patterns
Imagine the Future



Gartner's 2015 Hype Cycle for Emerging Technologies ([source](#))

Advanced Analytics
Hybrid Cloud
Internet of Things

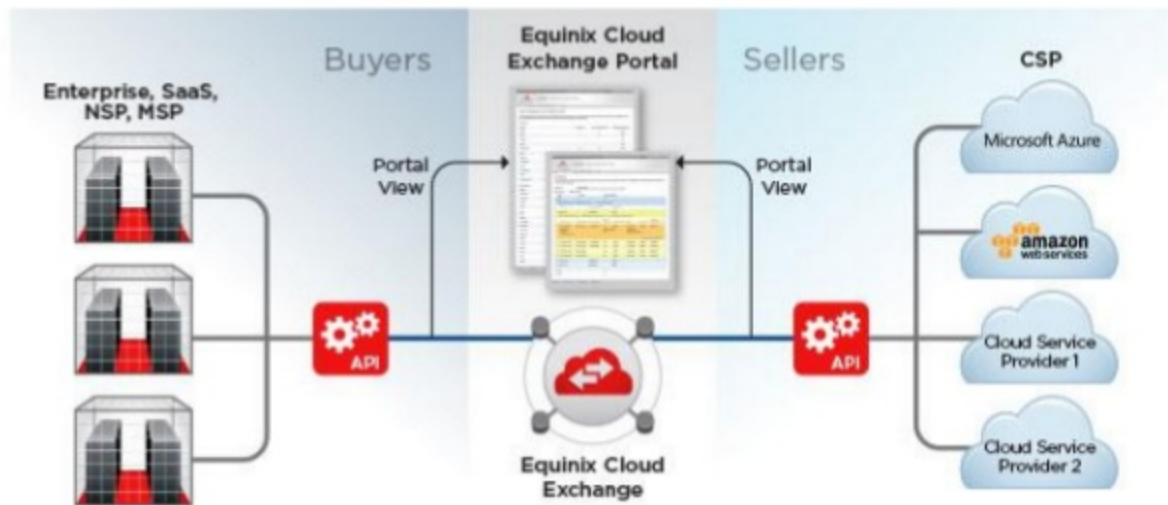


Hybrid Clouds

Cloud Exchange
Application Portability

Private-Public

Public-Public



Cloud Wars

AWS
Microsoft Azure
Google Cloud
VMWare
Openstack
etc.

cloudera



Data has Mass and
Gravity

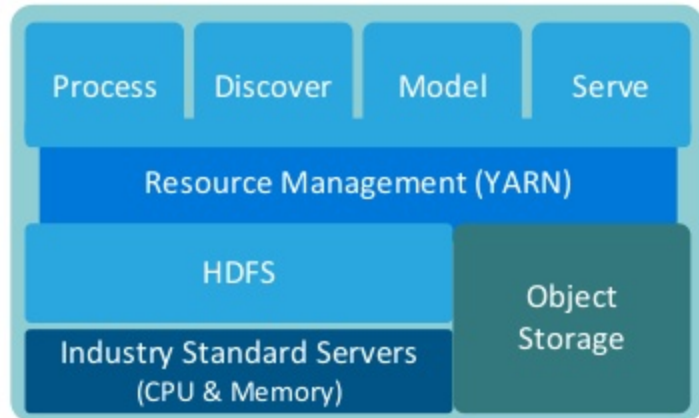
cloudera



Hadoop Environments

On-Premise versus Cloud

	On-Premise	Cloud
Storage	Direct Attached	Direct Attached or Object Store
Data	Not shared across clusters	Shared across multiple clusters
Sizing	Fixed-size	Dynamic based on load
Usage Model	All users share cluster	Clusters created as needed for apps/users



Cloud providers
shipping distributions
of Hadoop

Integration
Unlock Query Engines
Migration workloads

*Is that a sustainable
advantage? Or just a
temporary stop gap?*



Maturity level

On-prem vs. Cloud

Monitoring

Dev / Test / Prod

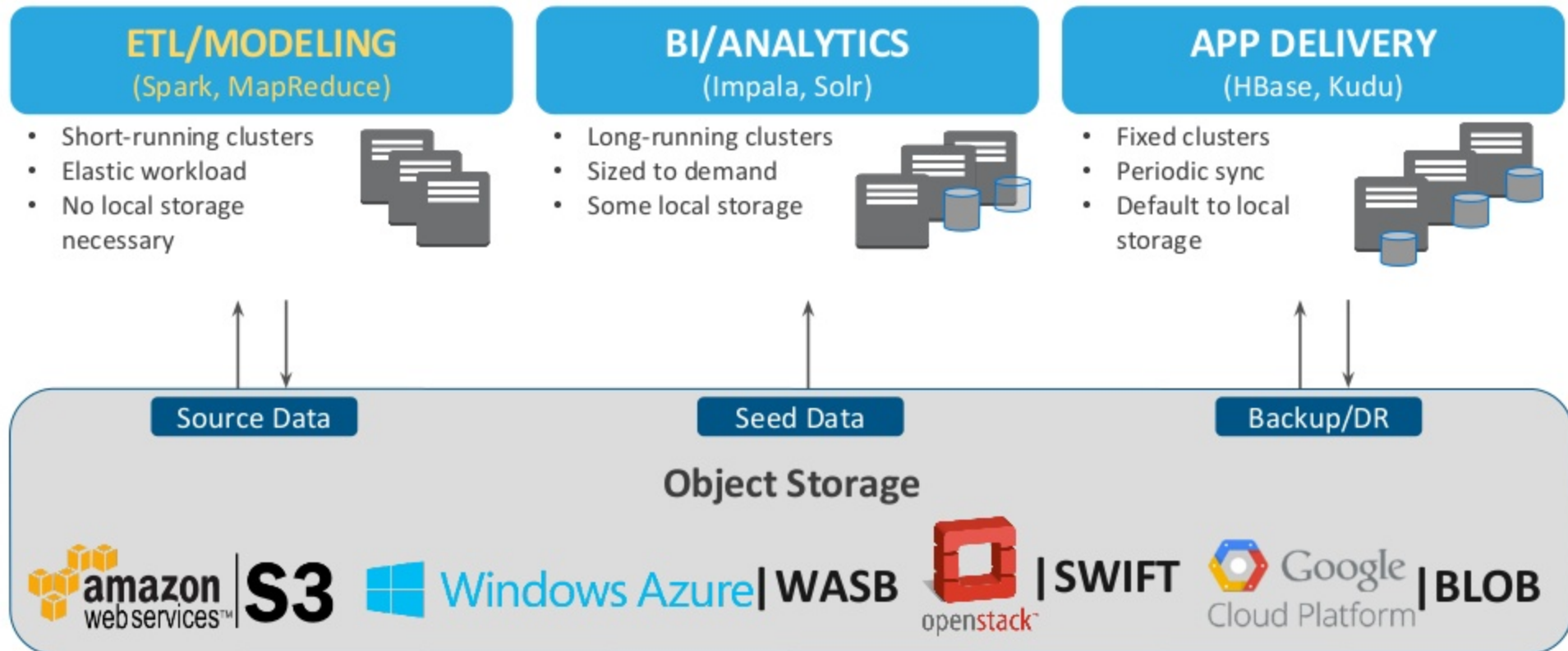
Availability

Durability

cloudera

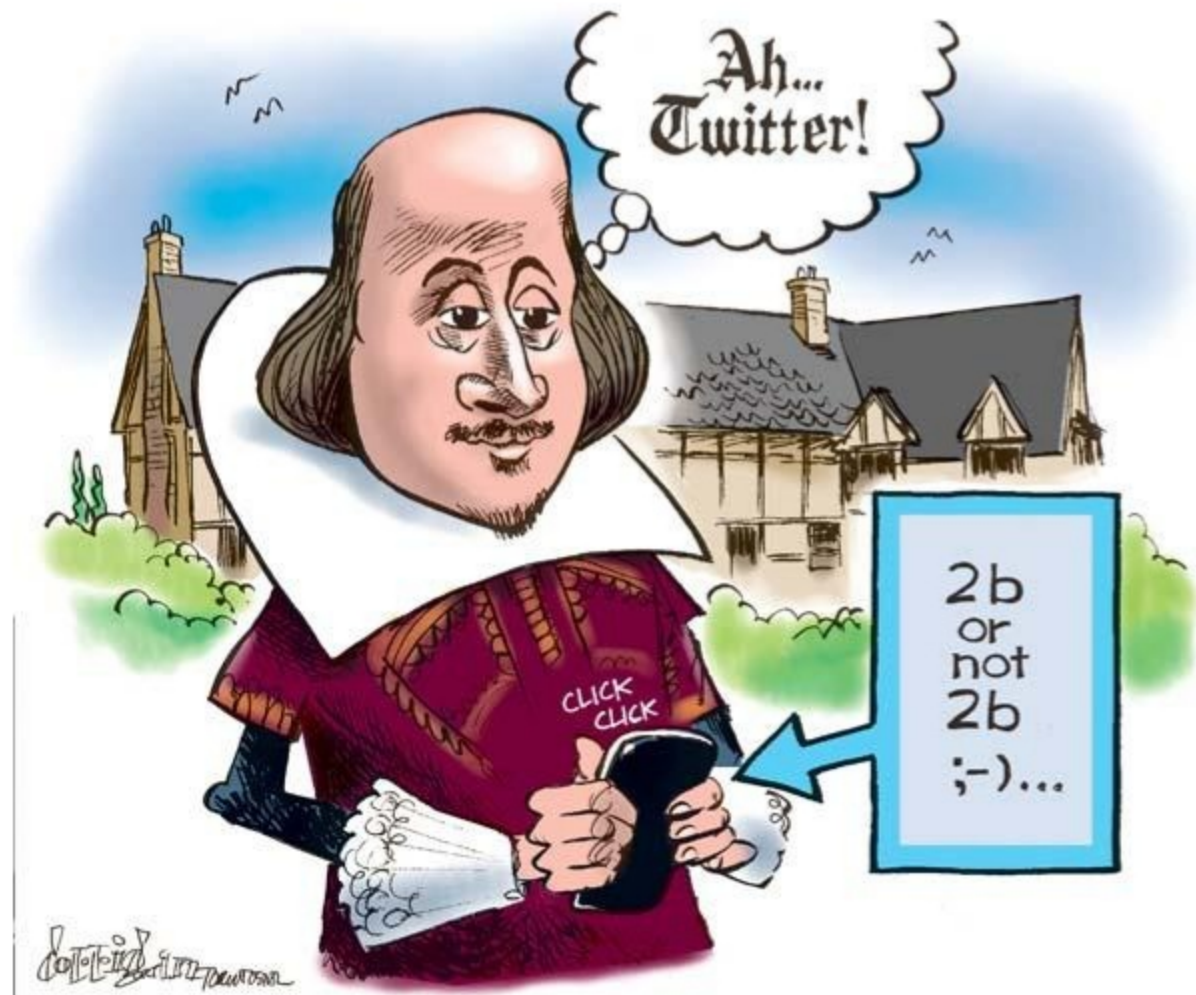


Common Architectural Patterns in the Cloud



Cluster lifecycle
management

Create / Terminate
Discovery
Metadata
Monitoring



Work Queue

Workflows

Dispatch

Tracking

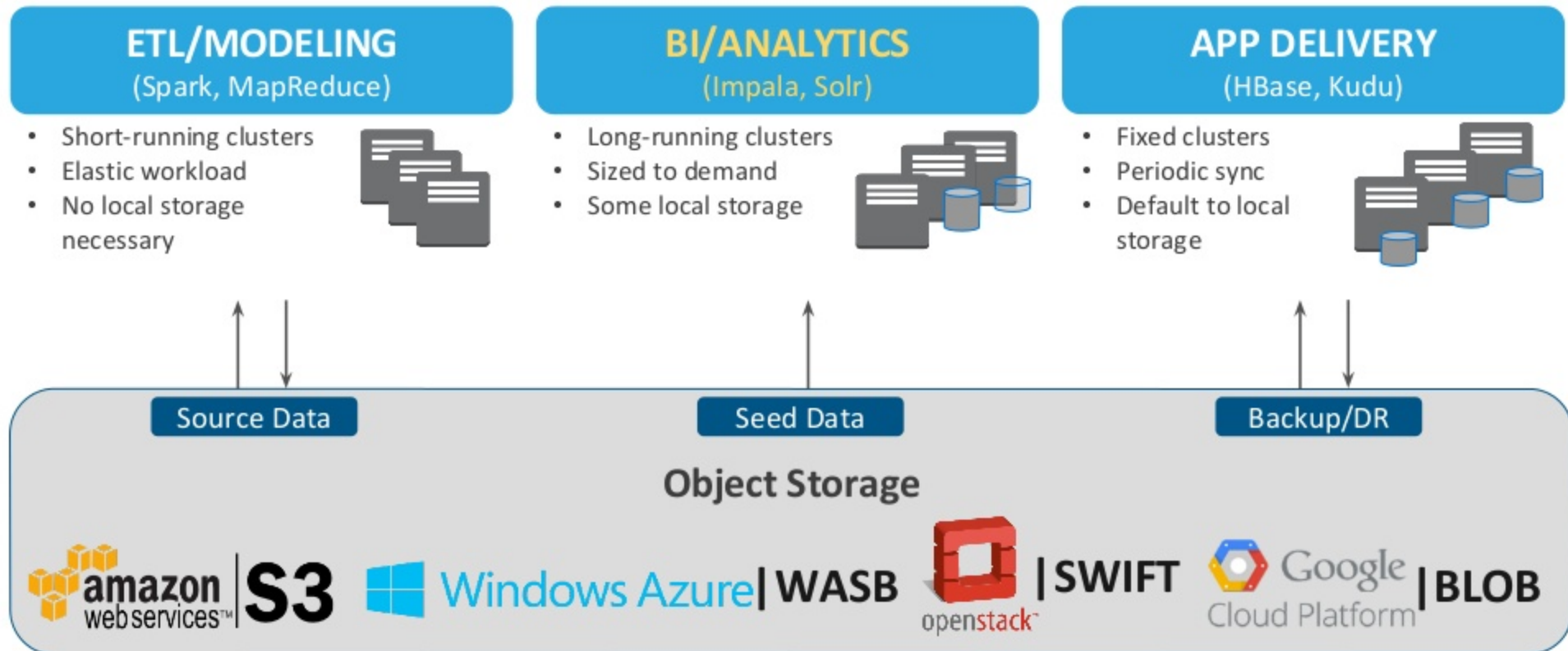
Decoupled

Fault Tolerant



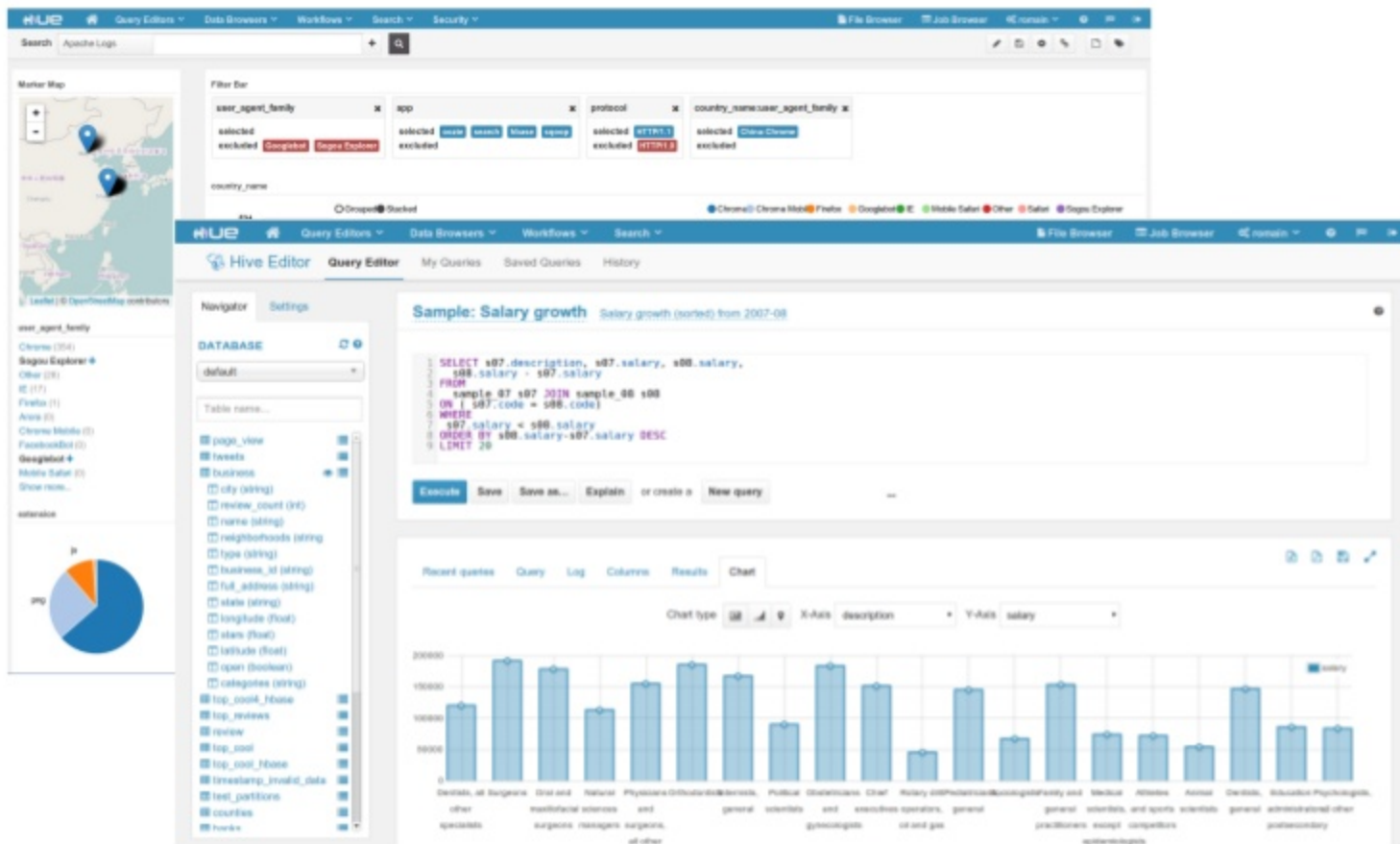
PLEASE FORM AN ORDERLY QUEUE

Common Architectural Patterns in the Cloud



Multi-user

Secure
Isolated
Friendly



Elastic

Grow or shrink

Business hours

Number of users

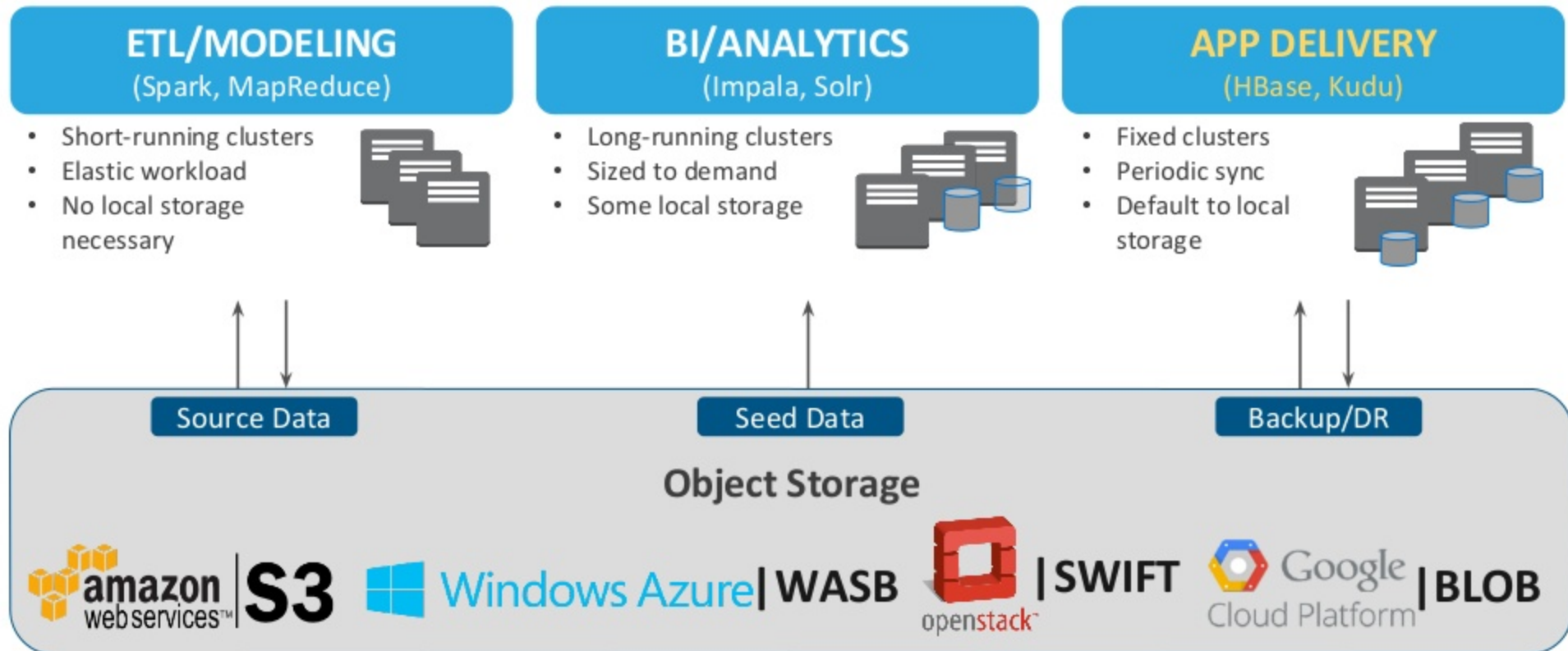
Storage vs. Compute

Cost efficient

cloudera



Common Architectural Patterns in the Cloud



Advanced Monitoring

Latency

Resource utilization

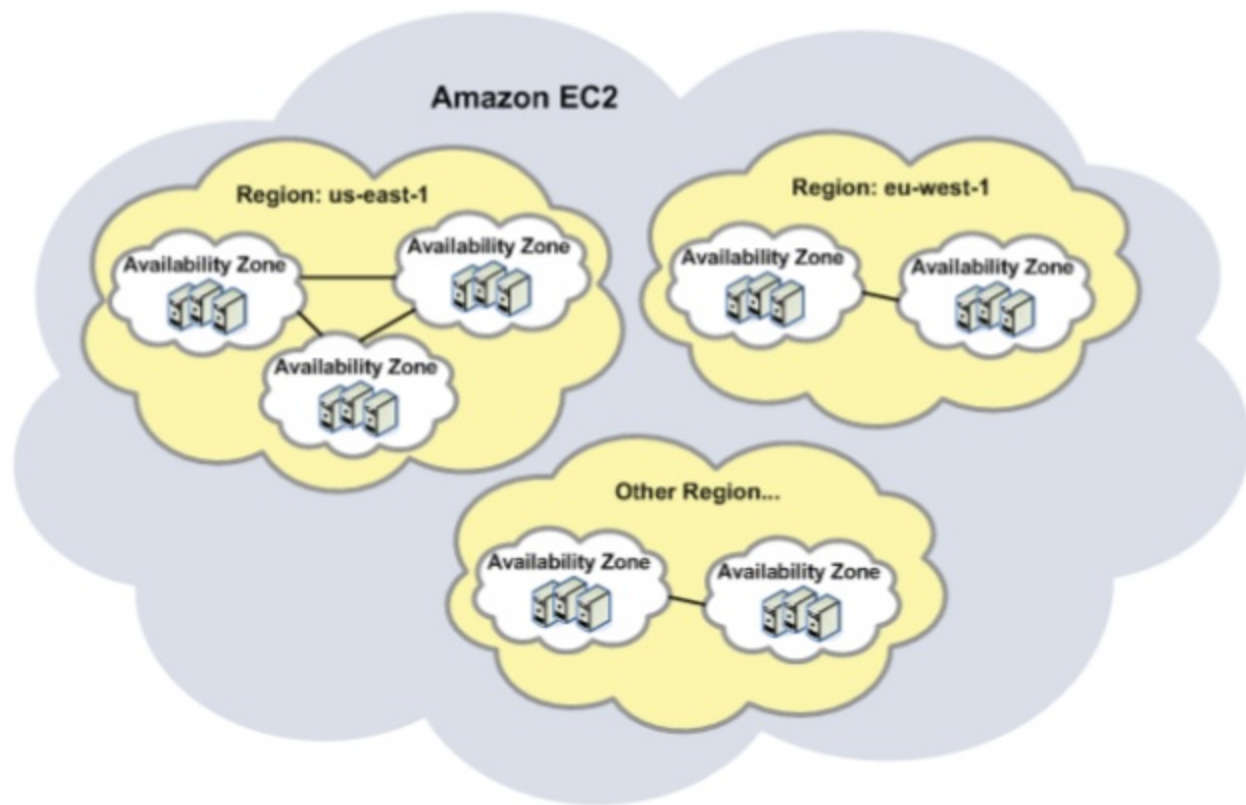
Consistent performance

cloudera



High availability and
failure domains

Data durability
Repair within SLA
Host-to-instance



Backup and disaster
recovery

Object store centric
Active-Standby

cloudera



Imagine the Future

Portable Experience

Self-service

Self-healing

Granular Security

Advanced Governance

Complete Management

What's your vision?

cloudera





- **Who:** Featuring data scientists from Salesforce, Uber, Pinterest, and more
- **When:** Thursday, October 22, 2015
- **Where:** Broadway Studios, San Francisco

Public registration is now open!

The conference for and by Data Scientists, from startup to enterprise

wrangleconf.com



cloudera

Thank you!

asavu@cloudera.com

Resources

Cloudera Director: <http://www.cloudera.com/director>

Interested in API level integration and scripting?

<https://github.com/cloudera/director-sdk>

<https://github.com/cloudera/director-scripts>

Interested in integration with another cloud platform?

<https://github.com/cloudera/director-spi>

<https://github.com/cloudera/director-google-plugin>

Resources

What's new in Cloudera Director 1.5?

<http://blog.cloudera.com/blog/2015/08/whats-new-in-cloudera-director-1-5/>

Get Started

[AWS Reference Guide](#)

[GCP Reference Guide](#)

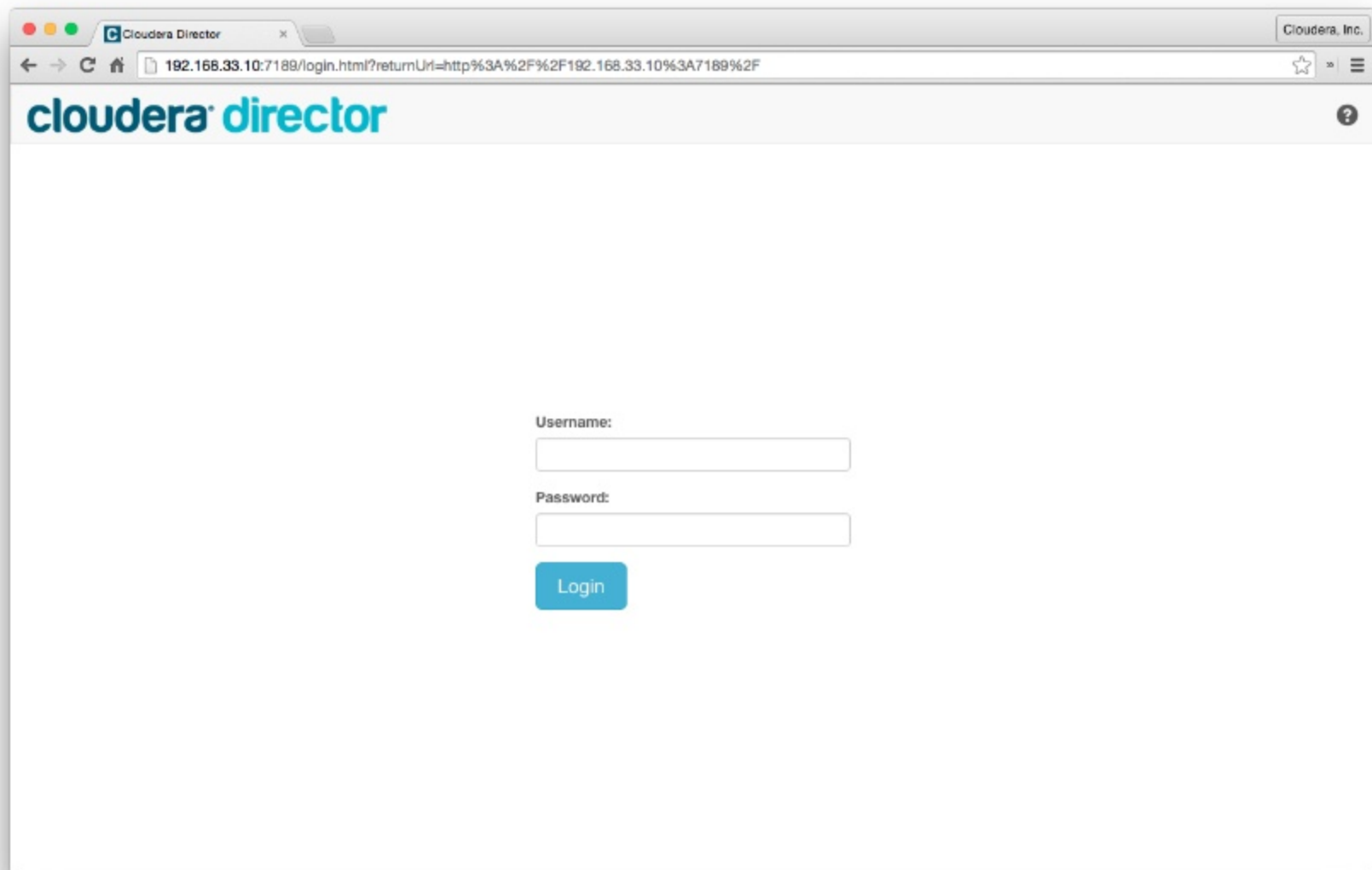
Try It Out

[AWS Quickstart](#)



cloudera

Cloudera Director
Screenshots



Welcome | Cloudera Director

Cloudera, Inc.

← → ↻ 🏠 📄 localhost:7189

☆ » ☰

cloudera director

admin ?

Welcome

This wizard will take you through the following steps to provision a cluster:

- Add Environment**
 - You will need the following:
 - Cloud Provider credentials.
 - SSH credentials (private key and username).
- Add Cloudera Manager**
 - When provisioning a new **Cloudera Manager** instance, you'll need to define an **Instance Template**. You will need the following:
 - Your virtual machine image ID.
 - Cloud Provider network settings.
 - A **Cloudera Manager** repository.
 - Optionally, you may define:
 - An **External Database** by providing a valid address (host and port) and credentials (username and password).
 - A **Database Server Template** for provisioning a database server from your Cloud Provider by providing credentials (username and password).
- Add Cluster**
 - You will need to define **Instance Groups**, each with:
 - An **Instance Template**.
 - A desired instance count.
 - Optionally, you may define an **External Database** or a **Database Server Template**.

Please see the [documentation](#) for more information!

Let's get started!

Cloudera, Inc.

← → ↻ 🏠 📄 localhost:7189/wizards/addEnvironment.html ☆ » ☰

cloudera director

admin ?

Add Environment

GENERAL INFORMATION

Environment name

?

Cloud provider

Amazon Web Services (AWS)

?

▼ Advanced Options

☒ Make credentials available as cluster client configurations

?

Access key ID

?

Secret access key

?

EC2

▼ Advanced Options

☒ Associate public IP addresses

?

IAM endpoint

?

EC2 region

us-east-1

?

⏪ Quit

▶▶ Continue

Cloudera, Inc.

localhost:7189/wizards/addEnvironment.html#step=st1

admin ?

cloudera director

Add Cloud

Env

CR

Database Se

Configuration

Quit

Continue

Instance Template

Instance Template name

?

Instance type

?

Image (AMI) ID

?

Tags

+

Security group IDs

-

+

?

VPC subnet ID

?

Advanced Options

Instance name prefix

director

?

IAM profile name

?

Placement group

?

Root volume size (GB)

50

?

Root volume type

gp2

?

SSH username

?

Bootstrap script

☒ File Upload

☐ Direct Input

?

Choose File

Cloudera, Inc.

localhost:7189/wizards/addEnvironment.html#step=st1

cloudera director

admin ?

Add Cloudera Environment

Database Server

Name ?

Master username ?

Master user password ?

DB engine ?

Tags

Allocated storage (GB) ?

Instance class ?

DB subnet group name ?

VPC security group IDs - + ?

▼ Advanced Options

Instance name prefix ?

☐ Auto minor version upgrade ?

Backup retention period (days) ?

DB name ?

DB parameter group name ?

◀ Quit

▶ Continue

Cloudera, Inc.

← → ↻ 🏠 localhost:7189/wizards/addEnvironment.html

cloudera director

admin ?

Add Environment

GENERAL INFORMATION

Environment name

?

Cloud provider

Google Cloud Platform

?

Project ID

?

▼ Advanced Options

☒ Make credentials available as cluster client configurations

?

Client ID JSON Key

☒ File Upload ☐ Direct Input

Choose File

?

GOOGLE COMPUTE ENGINE

▼ Advanced Options

Region

us-central1

?

⏪ Quit

▶ Continue

Cloudera, Inc.

localhost:7189/wizards/addEnvironment.html#step=st1

admin ?

cloudera director

Add Cloudera

Environment

Configuration

Database Setup

Quit

Continue

Instance Template

Instance Template name

?

Machine Type

?

Image Alias

?

Tags

+

Zone

?

Advanced Options

Instance name prefix

director

?

Network Name

default

?

Boot Disk Size (GB)

60

?

Data Disk Count

2

?

Data Disk Type

LocalSSD

?

Data Disk Size

375

?

Local SSD Interface Type

NVME

?

SSH username

?

Bootstrap script

File Upload

Direct Input

?

Cloudera Director

Cloudera, Inc.

192.168.33.10:7189/wizards/addEnvironment.html#step=st1

admin ?

cloudera director

Add Cloudera Manager

Environment AWS-1

Cloudera Manager name

CM-1 ?

☒ Enable Cloudera Enterprise trial ?

Instance Template

c42x ?

View ?

Repository

☐ Override default Cloudera Manager repository ?

Quit

Continue

Cloudera Director

Cloudera, Inc.

192.168.33.10:7189/wizards/addEnvironment.html#step=st2

admin ?

Add Cluster

Environment

AWS-1

Cloudera Manager

CM-1

Cluster name

CDH-1

?

Products

Name

Version

CDH

5

?

Parcel Repositories

☐ Override default

?

Services

☐ Core Hadoop

HDFS, Hive, Hue, Oozie, Sqoop 2, YARN, ZooKeeper

☐ Core Hadoop with HBase

HBase, HDFS, Hive, Hue, Oozie, Sqoop 2, YARN, ZooKeeper

☐ Core Hadoop with Impala

HDFS, Hive, Hue, Impala, Oozie, Sqoop 2, YARN, ZooKeeper

☐ Core Hadoop with Search

HDFS, Hive, Hue, Oozie, Solr, Sqoop 2, YARN, ZooKeeper

Quit

Continue

Cloudera Director

Cloudera, Inc.

192.168.33.10:7189/wizards/addEnvironment.html#step=st2

Services

☐ Core Hadoop

HDFS, Hive, Hue, Oozie, Sqoop 2, YARN, ZooKeeper

☐ Core Hadoop with HBase

HBase, HDFS, Hive, Hue, Oozie, Sqoop 2, YARN, ZooKeeper

☐ Core Hadoop with Impala

HDFS, Hive, Hue, Impala, Oozie, Sqoop 2, YARN, ZooKeeper

☐ Core Hadoop with Search

HDFS, Hive, Hue, Oozie, Solr, Sqoop 2, YARN, ZooKeeper

☒ Core Hadoop with Spark on YARN

HDFS, Hive, Hue, Oozie, Spark on YARN, Sqoop 2, YARN, ZooKeeper

☐ All Services

Flume, HBase, HDFS, Hive, Hue, Impala, Key-Value Store Indexer, Oozie, Solr, Spark on YARN, Sqoop 2, YARN, ZooKeeper

Instance groups

Quit

Continue

Cloudera Director x Cloudera, Inc.

192.168.33.10:7189/wizards/addEnvironment.html#step=st3

cloudera director admin ?

Status

CM-1 Bootstrapping

9 / 21

REQUESTING AN INSTANCE FOR CLOUDERA MANAGER

- Starting

CDH-1 Bootstrapping

6 / 26

REQUESTING 4 INSTANCE(S) IN 2 GROUP(S)

- Starting
- Starting

Quit Continue

Modify Cluster - CDH-1

Cloudera, Inc.

nightly-1.lpad.cloudera.com:7189/wizards/growAndShrink.html?environment=C5+Test+Environment&deployment=CM-1&cluster=CDH-1

admin ?

cloudera director

Modify Cluster

Cluster name

CDH-1

Products

Name	Version
CDH	5

Services

HDFS, Hive, Hue, Oozie, Spark on YARN, Sqoop 2, YARN, ZooKeeper

Instance groups

Name	Roles	Instance Template	Instance Count	Min Count ?
masters	View Roles	View Template	1 Edit	<input type="text" value="1"/> Delete Group
workers	View Roles	View Template	New total: 8 Edit	<input type="text" value="3"/> Delete Group
		To be added:	5	
		To be repaired:	1	
Add Group				

[Quit](#)[Continue](#)

192.168.33.10:7189/wizard x Cloudera Director x Cloudera Director API x Cloudera, Inc.

localhost:7189/api-console/index.html#/environments

Cloudera Director

http://localhost:7189/api-docs?group=v2

Explore

API Console

Software-defined EDH clusters on cloud infrastructure

[Terms of service](#)
[Contact the developer](#)
[Commercial License \(c\) 2014 Cloudera Inc.](#)

authentication : Authentication

Show/Hide | List Operations | Expand Operations | Raw

clusters : Clusters

Show/Hide | List Operations | Expand Operations | Raw

database-servers : Database Servers

Show/Hide | List Operations | Expand Operations | Raw

deployments : Deployments

Show/Hide | List Operations | Expand Operations | Raw

environments : Environments

Show/Hide | List Operations | Expand Operations | Raw

POST

/api/v2/environments

Create a new environment

GET

/api/v2/environments

List all environments

PUT

/api/v2/environments/{name}

Update an existing environment

GET

/api/v2/environments/{name}

Get an environment by name

DELETE

/api/v2/environments/{name}

Delete an environment by name

instance-templates : Instance Templates

Show/Hide | List Operations | Expand Operations | Raw

users : Users

Show/Hide | List Operations | Expand Operations | Raw

[BASE URL: http://localhost:7189/api-docs - API VERSION: 2.0]

192.168.33.10:7189/wizard

Cloudera Director

Cloudera Director API

Cloudera, Inc.

localhost:7189/api-console/index.html#/environments/list

401 Unauthorized

404 Not Found

Try it out! [Hide Response](#)

Request URL

http://localhost:7189/api/v2/environments

Response Body

```
[
  "AMS-1"
]
```

Response Code

200

Response Headers

```
{
  "Pragma": "no-cache",
  "X-Content-Type-Options": "nosniff",
  "Server": "Jetty(8.1.15.v20140411)",
  "X-Frame-Options": "DENY",
  "Transfer-Encoding": "chunked",
  "Content-Type": "application/json; charset=UTF-8",
  "Cache-Control": "no-cache, no-store, max-age=0, must-revalidate",
  "X-Xss-Protection": "1; mode=block",
  "X-Application-Context": "application:production:7189",
  "Expires": "0"
}
```

PUT

/api/v2/environments/{name}

Update an existing environment

GET

/api/v2/environments/{name}

Get an environment by name

DELETE

/api/v2/environments/{name}

Delete an environment by name

192.168.33.10:7189/wizard x Cloudera Director x cloudera/director-sdk x Cloudera, Inc.

← → ↺ 🏠 GitHub, Inc. [US] https://github.com/cloudera/director-sdk ☆ » ☰

🔍 This repository Search Pull requests Issues Gist + 👤

cloudera / director-sdk Unwatch 24 ★ Star 3 🍴 Fork 1

Cloudera Director API clients <http://www.cloudera.com/director> — Edit

📦 5 commits 🌿 1 branch 📦 0 releases 👤 3 contributors

🔄 Branch: master director-sdk / + ⋮

Updated legal information in readme.

👤 jadair-cloudera authored on Jun 15 latest commit 19755d3809

📁 java-client-samples	Add Python and Java examples for cluster creation	5 months ago
📁 java-client	Import Cloudera Director 1.1 API clients for Java and Python (Includi...	5 months ago
📁 python-client-samples	Add Python and Java examples for cluster creation	5 months ago
📁 python-client	Import Cloudera Director 1.1 API clients for Java and Python (Includi...	5 months ago
📄 .gitignore	Import Cloudera Director 1.1 API clients for Java and Python (Includi...	5 months ago
📄 README.md	Updated legal information in readme.	a month ago

📄 README.md

Cloudera Director API clients

Cloudera Director is the reliable, self-service experience for Cloudera's enterprise data hub in the cloud - allowing enterprises to unlock all the benefits of Hadoop in the cloud, without compromising on the critical aspects of an enterprise-grade, Hadoop platform.

🔗 Code

🔍 Issues 0

🔍 Pull requests 0

📖 Wiki

📊 Pulse

📊 Graphs

⚙ Settings

SSH clone URL

git@github.com:clc

You can clone with HTTPS, SSH, or Subversion.

📄 Clone in Desktop

📄 Download ZIP



cloudera

Thank you!

asavu@cloudera.com