Final Takeaway

Xavier Morera

HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera   www.xaviermorera.com

Final Takeaway

Why Spark with Scala & Cloudera?

- History

- What we will cover in this training

- What do you need for this course?

Final Takeaway

Getting an Environment & Data

- Cloudera cluster for Spark on Yarn or Spark Standalone
- StackOverflow / StackExchange

Final Takeaway

Scala Fundamentals for This Course
- History, philosophy and paradigm
  - Functional programming
- Scala
  - Interactive shell: REPL
  - Use scalac/scala
  - sbt
- Covered the basics
  - Strongly typed
- spark2-shell: Spark in the shell

Final Takeaway

Understanding Spark: An overview

- Coarse grained transformations
- Parallelism and pipelining
- Narrow and wide transformations
- Lazy execution and lineage
- Libraries

Final Takeaway

Getting Technical with Spark

- Storage and file formats
- Spark APIs
- Performance, configuration & logging
- Visualizing your Spark application
- Navigating the documentation

Final Takeaway

Learning the Core of Spark: RDDs

- SparkContext
- Resilient Distributed Datasets
- Many ways of creating RDDs
- Return data to the Driver
- Save RDDs

Final Takeaway

Going Deeper into Spark Core

- Anonymous functions
- Transformations
- Actions
- Partition operations
- Set operations and aggregations
- Caching and data persistence
- Shared variables
- What's better?

Final Takeaway

Increasing Proficiency with Spark: DataFrames & Spark SQL

- "Everyone uses SQL"
- SparkSession
- Creating DataFrames
- Schemas
- Rows, columns, expressions & operators
- UDFs

Final Takeaway

Continuing the Journey with DataFrames & Spark SQL

- DataFrame DSL
- Saving DataFrames
- Spark SQL & Temporary Views
- Peristent Tables
- Hive support & external databases
- Aggregating, grouping & joining
- Catalog API

Final Takeaway

Understanding a Typed API: Datasets

- Higher-level typed API
- DataFrames = Dataset[Row]
- Typed and untyped operations
- Functional programming

Continuing the Journey with Spark, Scala & Cloudera: Next Steps

Getting a Hadoop Cluster

CDH + Tools
On-prem & Cloud

Director
Cloud

Altus
Platform-as-a-Service

Cloudera Altus

Platform-as-a-Service
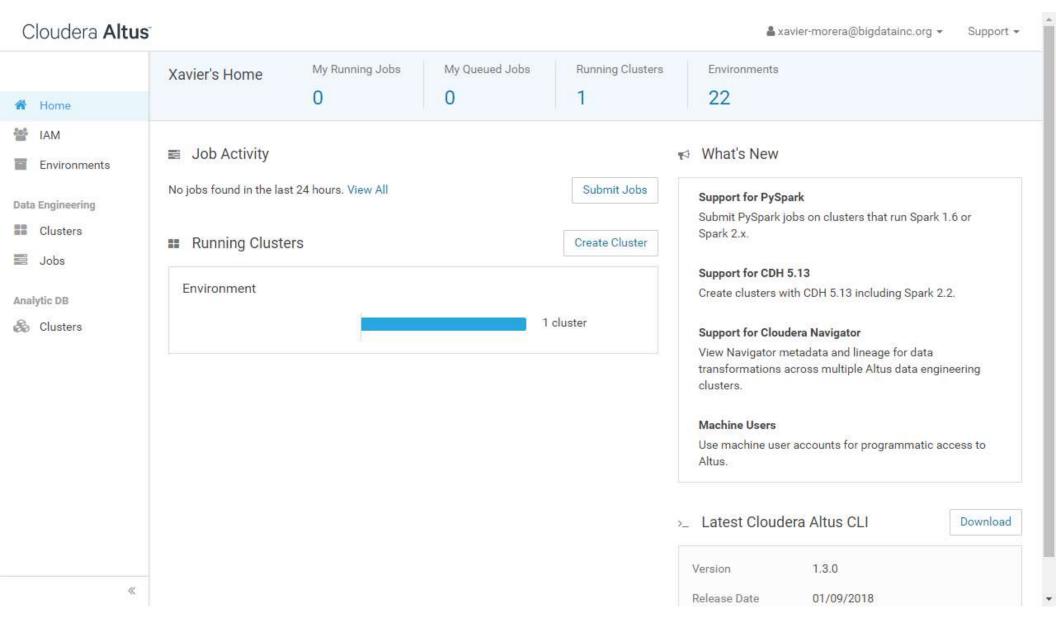
Provision clusters in the cloud
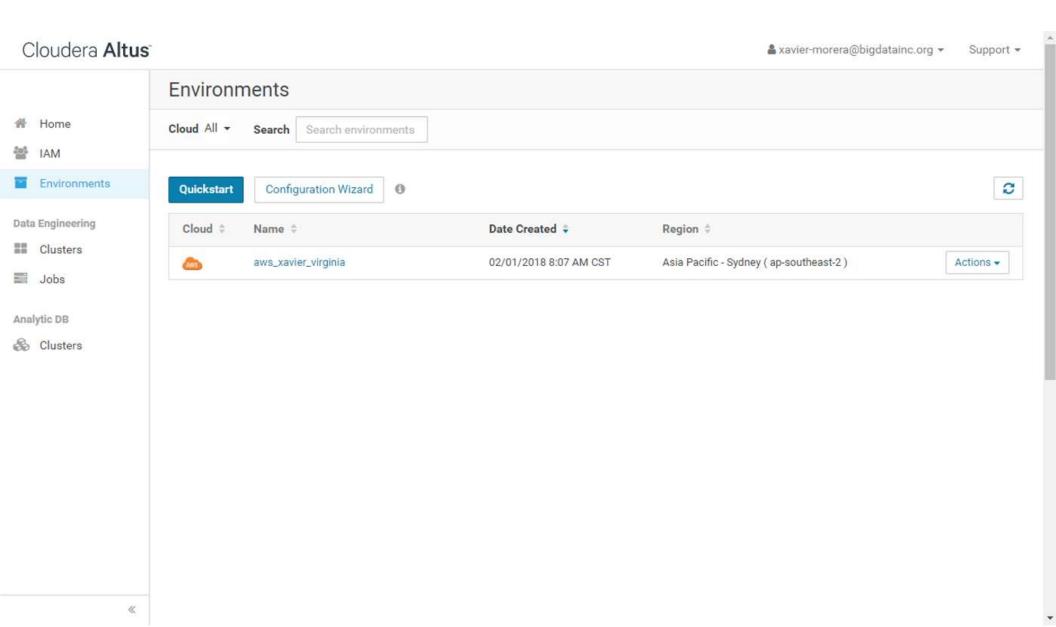- AWS & Azure
- Managed cluster

Data Engineering & Workload Analytics
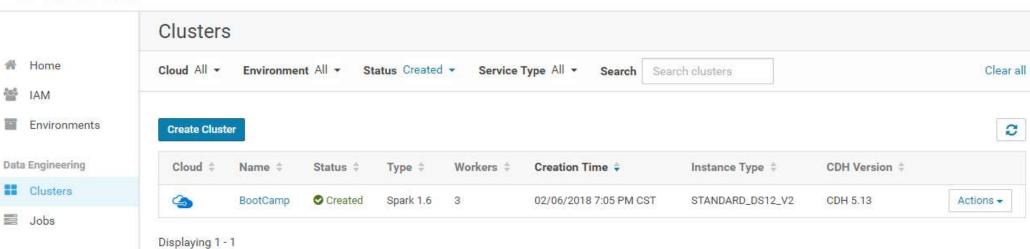
Processing engines
- MR2, Hive on MR2, Spark & Hive on Spark

Cloudera Altus

Xavier's Home

My Running Jobs	My Queued Jobs	Running Clusters	Environments
0	0	1	22

🏠 **Home**

👥 IAM

🗄 Environments

Data Engineering

▦ Clusters

☰ Jobs

Analytic DB

🎱 Clusters

☰ **Job Activity**

No jobs found in the last 24 hours. View All   **Submit Jobs**

▦ **Running Clusters**   **Create Cluster**

Environment

                                                    1 cluster

📢 **What's New**

Support for PySpark

Submit PySpark jobs on clusters that run Spark 1.6 or Spark 2.x.

Support for CDH 5.13

Create clusters with CDH 5.13 including Spark 2.2.

Support for Cloudera Navigator

View Navigator metadata and lineage for data transformations across multiple Altus data engineering clusters.

Machine Users

Use machine user accounts for programmatic access to Altus.

>_ **Latest Cloudera Altus CLI**   **Download**

Version	1.3.0
Release Date	01/09/2018

«

Cloudera Altus

Environments

Cloud All ▾   Search [Search environments]

Quickstart   Configuration Wizard   ⓘ   ⟳

Cloud ⇕	Name ⇕	Date Created ⇕	Region ⇕	
AWS	aws_xavier_virginia	02/01/2018 8:07 AM CST	Asia Pacific - Sydney (ap-southeast-2)	Actions ▾

Home
IAM
Environments

Data Engineering
Clusters
Jobs

Analytic DB
Clusters

«

Cloudera **Altus**™

Home
IAM
Environments

Data Engineering
Clusters
Jobs

Analytic DB
Clusters

Clusters

Cloud All ▾   Environment All ▾   Status Created ▾   Service Type All ▾   Search [Search clusters]   Clear all

Create Cluster   🔄

Cloud ⇕	Name ⇕	Status ⇕	Type ⇕	Workers ⇕	Creation Time ⇕	Instance Type ⇕	CDH Version ⇕	
☁	BootCamp	✔ Created	Spark 1.6	3	02/06/2018 7:05 PM CST	STANDARD_DS12_V2	CDH 5.13	Actions ▾

Displaying 1 - 1

«

Submit Jobs

Job Settings

Submission *   ⦿ Single job   ⦾ Group of jobs

Job Type * ⓘ

Select job type ▾
Select job type
Hive
MapReduce2
PySpark
Spark

Cancel   **Submit Jobs**

Submit Jobs

Submission *	⦿ Single job ○ Group of jobs
Job Type * ⓘ	Spark ▾
Job Name ⓘ	Enter a name for this job
Main Python File * ⓘ	path/to/main.py
Python File Dependencies ⓘ	path/to/file with py, zip, or egg extension ⊟ ⊞
Application Arguments ⓘ	Enter application arguments ⊟ ⊞
Spark Arguments ⓘ	☐
Action on Failure * ⓘ	None ▾

Cluster Settings

Cluster * ⓘ	⦿ Use existing ○ Create new ○ Clone existing
	Select cluster ▾
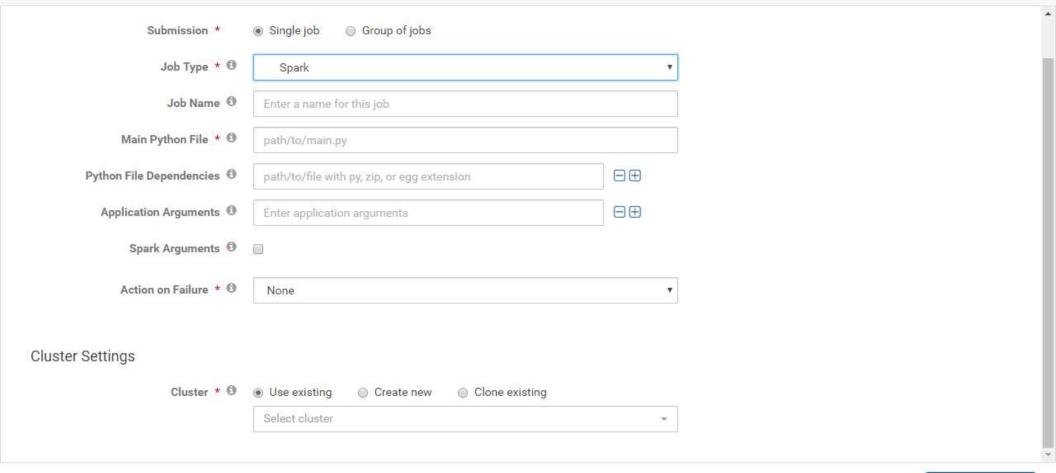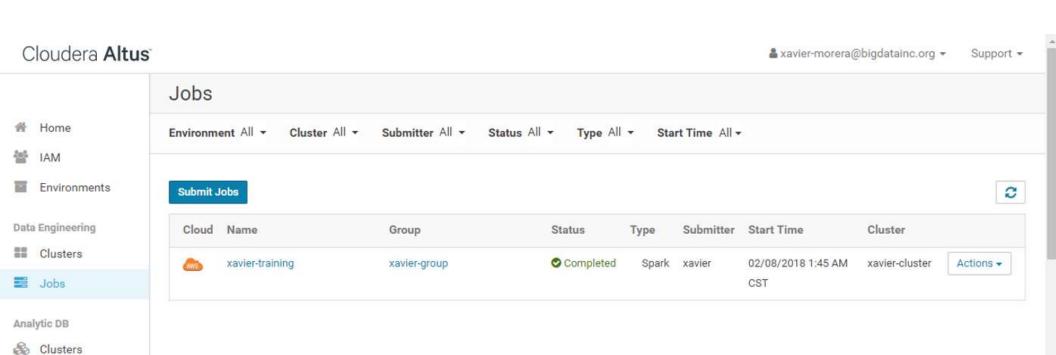
Cancel   **Submit Jobs**

Cloudera Altus™

Home
IAM
Environments

Data Engineering
- Clusters
- Jobs

Analytic DB
- Clusters

Jobs

Environment All ▾   Cluster All ▾   Submitter All ▾   Status All ▾   Type All ▾   Start Time All ▾

Submit Jobs   ⟳

Cloud	Name	Group	Status	Type	Submitter	Start Time	Cluster	
AWS	xavier-training	xavier-group	✅ Completed	Spark	xavier	02/08/2018 1:45 AM CST	xavier-cluster	Actions ▾

Cloudera Data Science Workbench

Web application for working with Big Data

- Data Science
- From exploration to production

Work with R, Python & Scala

Self-Service Data Science for the Enterprise

Installing and Upgrading Cloudera Data Science Workbench 1.3.x Using Cloudera Manager

This topic describes how to install and upgrade Cloudera Data Science Workbench using Cloudera Manager.

- Installing Cloudera Data Science Workbench 1.3.x Using Cloudera Manager
- Upgrading to the Latest Version of Cloudera Data Science Workbench 1.3.x

Installing Cloudera Data Science Workbench 1.3.x Using Cloudera Manager

Use the following steps to install Cloudera Data Science Workbench using Cloudera Manager.

1. Prerequisites
2. Install Cloudera Distribution of Apache Spark 2
3. Configure JAVA_HOME (Required for Spark 2.2)
4. Download and Install the Cloudera Data Science Workbench CSD
5. Install the Cloudera Data Science Workbench Parcel
6. Add the Cloudera Data Science Workbench Service
7. Create the Administrator Account
8. Next Steps

Prerequisites
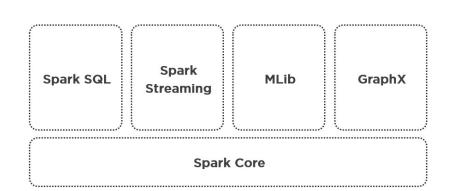
Before you begin installing Cloudera Data Science Workbench, make sure you have completed the steps to

http://tiny.bigdatainc.org/getcdsw

Streaming

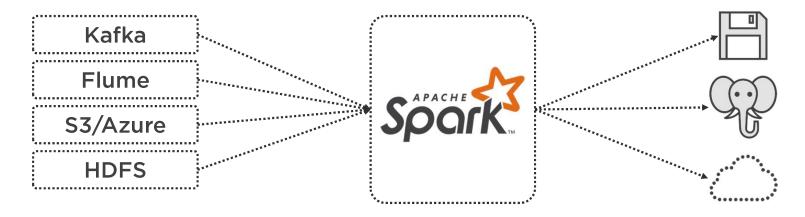Process live data streams in a scalable, high-throughput and fault tolerant way

Spark Streaming

Spark SQL	Spark Streaming	MLib	GraphX

Spark Core

Process live data streams in a scalable, high-throughput and fault tolerant way using Apache Spark

Spark Streaming

Process live data streams in a scalable, high-throughput and fault tolerant way using Apache Spark

Structured Streaming

Process live data streams in a scalable, high-throughput and fault tolerant way using Apache Spark

TensorFlow

Open-source library

- Created by Google

Machine learning applications

- Deep learning

Use with Spark?

TensorFrames

TensorFlow on Spark DataFrames

Manipulate DataFrames
- TensorFlow programs

Machine Learning

tensorframes (homepage)

Tensorflow wrapper for DataFrames on Apache Spark

@databricks / ⭐⭐⭐⭐⭐ (👤3)

TensorFrames (TensorFlow on Spark Dataframes) lets you manipulate Spark's DataFrames with TensorFlow programs.
This package provides Python bindings, a Scala DSL and a small runtime to express and run TensorFlow computation graphs.

Tags

2 tensorflow

How to [+]

Include this package in your Spark Applications using:

spark-shell, pyspark, or spark-submit

```
> $SPARK_HOME/bin/spark-shell --packages databricks:tensorframes:0.2.9-s_2.11
```

Releases

Version: 0.2.9-s_2.11 (4be3b2 | zip | jar) / Date: 2017-09-05 / License: Apache-2.0

Version: 0.2.9-s_2.10 (4be3b2 | zip | jar) / Date: 2017-09-05 / License: Apache-2.0

Version: 0.2.9-rc3-s_2.11 (99e6d8 | zip | jar) / Date: 2017-08-31 / License: Apache-2.0

Version: 0.2.9-rc3-s_2.10 (99e6d8 | zip | jar) / Date: 2017-08-31 / License: Apache-2.0

Thank You!

Xavier Morera

HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera   www.xaviermorera.com