H2O Rains With Databricks Cloud for Spark

Michal Malohlava <michal@h2o.ai>, H2O

Richard Garris <richard@databricks.com>, Databricks

H2O WORLD 2015

Spark

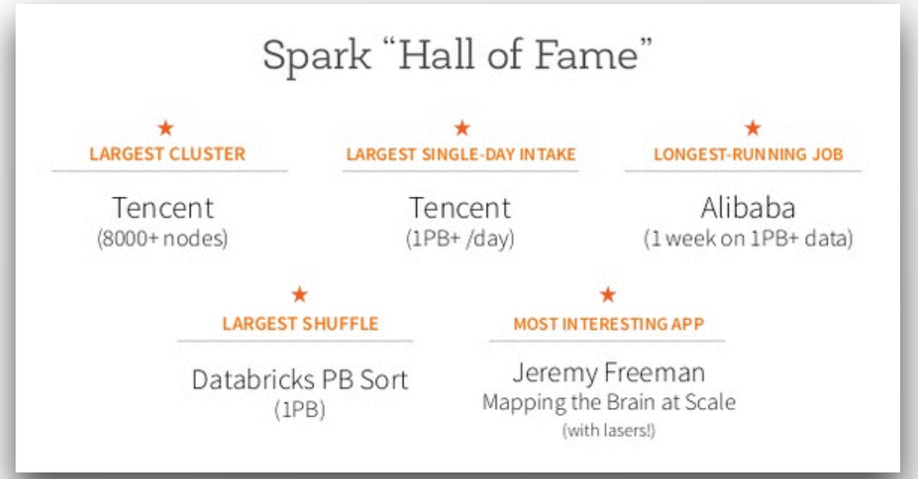
Open-source distributed execution platform

User-friendly API for data transformation based on RDDs, DataFrames and Datasets

Platform components - SparkSQL, MLLib, Streaming, GraphX

Multitenancy

Large and active community



Databricks

- Databricks
  - founded by the creators of Apache Spark
  - still contribute 75% of the code to the Spark project
  - cloud platform for running Spark in your AWS account

- Databricks Platform
  - integrated collaborative data science workspace
  - notebook interface inspired by iPython and Zeplin but purpose built for Spark
  - self service cluster manager and job scheduler for production Spark workloads

Can I run H2O on top of Databricks cloud?

YES,
You can!

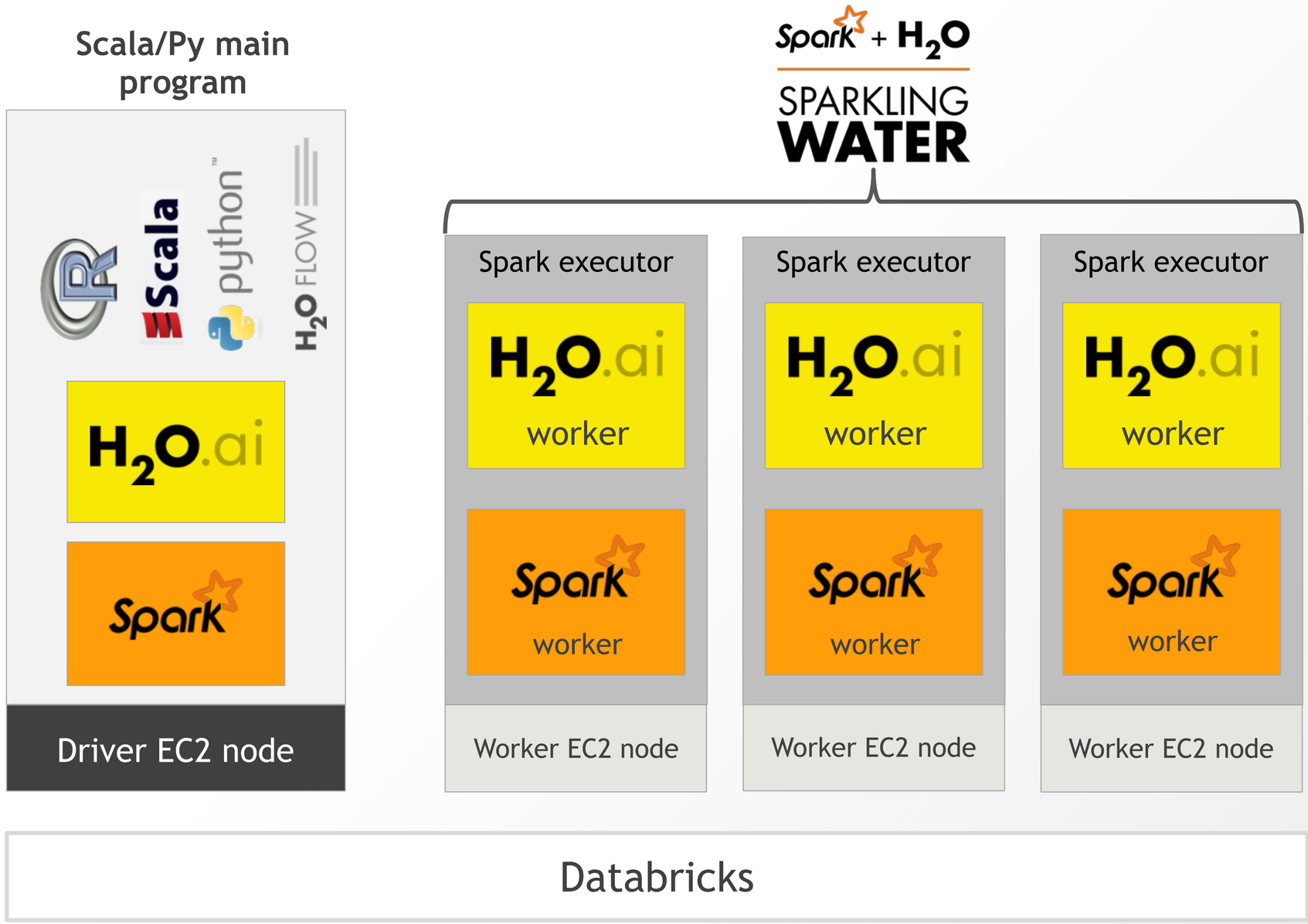
Sparkling Water

Provides

- Transparent integration of H2O with Spark ecosystem
- Transparent use of H2O data structures and algorithms with Spark API
- Platform for building Smarter Applications

Excels in existing Spark workflows requiring advanced Machine Learning algorithms

Databricks with H2O



Lets play with it!

What do you need?

- Databricks account (14 day free trial at www.databricks.com)
- Your AWS account
- Sparkling Water jar
- And some cool machine learning idea!

OR

Detect spam text messages

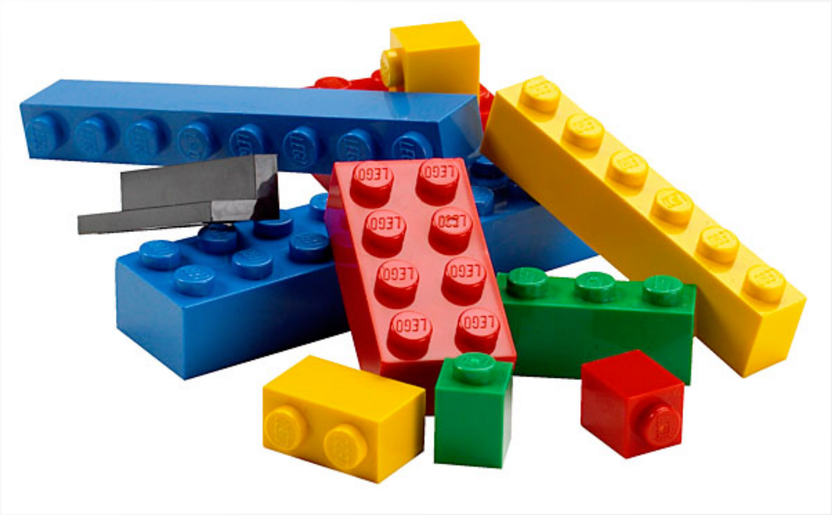
Data sample

	A	B
1	ham	Ok... But they said i've got wisdom teeth hidden inside n mayb need 2 remove.
2	ham	U thk of wat to eat tonight.
3	ham	I dunno until when... Lets go learn pilates...
4	spam	Someonone you know is trying to contact you via our dating service! To find out who it could be call from your mobile or landline 09064015307 BOX334SK38ch
5	ham	Ok c u then.
6	spam	URGENT! We are trying to contact U. Todays draw shows that you have won a £800 prize GUARANTEED. Call 09050003091 from land line. Claim C52. Valid12hrs only
7	spam	Not heard from U4 a while. Call 4 rude chat private line 01223585334 to cum. Wan 2C pics of me gettin shagged then text PIX to 8552. 2End send STOP 8552 SAM xxx
8	ham	staff.science.nus.edu.sg/~phyhcmk/teaching/pc1323
9	ham	Thank god they are in bed!
10	ham	Hey tmr meet at bugis 930 ?
11	spam	You are a winner you have been specially selected to receive £1000 cash or a £2000 award. Speak to a live operator to claim call 087123002209am-7pm. Cost 10p
12	spam	URGENT! Your Mobile No. was awarded £2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM
13	spam	Loan for any purpose £500 - £75,000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1956669 or text back 'help'
14	ham	Haha... Sounds crazy, dunno can tahan anot...
15	spam	You have won ?spam 000 cash or a ?2,000 prize! To claim, call09050000327
16	ham	Sorry i din lock my keypad.
17	ham	Thanx but my birthday is over already.
18	spam	FREE for 1st week! No1 Nokia tone 4 ur mobile every week just txt NOKIA to 8077 Get txting and tell ur mates. www.getzed.co.uk POBox 36504 W45WQ 16+ norm150p/tone
19	spam	Congratulations - Thanks to a good friend U have WON the £2,000 Xmas prize. 2 claim is easy, just call 08712103738 NOW! Only 10p per minute. BT-national-rate
20	ham	Me n him so funny...
21	spam	pdate_Now - Double mins and 1000 txts on Orange tariffs. Latest Motorola, SonyEricsson & Nokia & Bluetooth FREE! Call MobileUpd8 on 08000839402 or call2optout/!YHL
22	ham	Ok...
23	ham	Yup no more already... Thanx 4 printing n handing it up.
24	ham	Anything lor. Juz both of us lor.
25	ham	It's é only $140 ard...É rest all ard $180 at least...Which is é price 4 é 2 bedrm ($900)
26	ham	Oh oh... Den muz change plan liao... Go back have to yan jiu again...
27	ham	Ok lor then we go tog lor...
28	ham	Okay lor... Wah... like that def they wont let us go... Haha... What did they say in the terms and conditions?
29	ham	Dunno lei... I thk mum lazy to go out... I neva ask her yet...
30	ham	THATS ALRITE GIRL, U KNOW GAIL IS NEVA WRONG!!TAKE CARE SWEET AND DONT WORRY.C U L8TR HUN!LOVE Yaxxx

Machine Learning Workflow

Goal: For a given text message identify if it is spam or not

1. Extract data
2. Transform, tokenize messages
3. Build Tf-IDF model
4. Create and evaluate Deep Learning model
5. Use the model to detect spam



Databricks setup step-by-step

- Create a cluster
- Setup a library
  - Maven coordinate ai.h2o:sparkling-water-core_2.10:1.5.6
  - Attach library to cluster
- Load data
  - upload and create a table
- Create a new notebook
- Expose driver's URL
  - assign elastic IP or create a proxy
- Start coding

Live demo!

H2O.ai

Thank you!

Learn more at h2o.ai
Follow us at @h2oai