How Companies are Using Spark

And where the Edge in Big Data will be

Matei Zaharia

DATABRICKS

History

Decreasing storage costs have led to an explosion of big data

Commodity cluster software, like Hadoop, has made it 10-20x cheaper to store large datasets

Broadly available from multiple vendors

Implication

Big data storage is becoming commoditized, so how will organizations get an edge?

What matters now is what you can *do* with the data.

Two Factors

Speed: how quickly can you go from data to decisions?

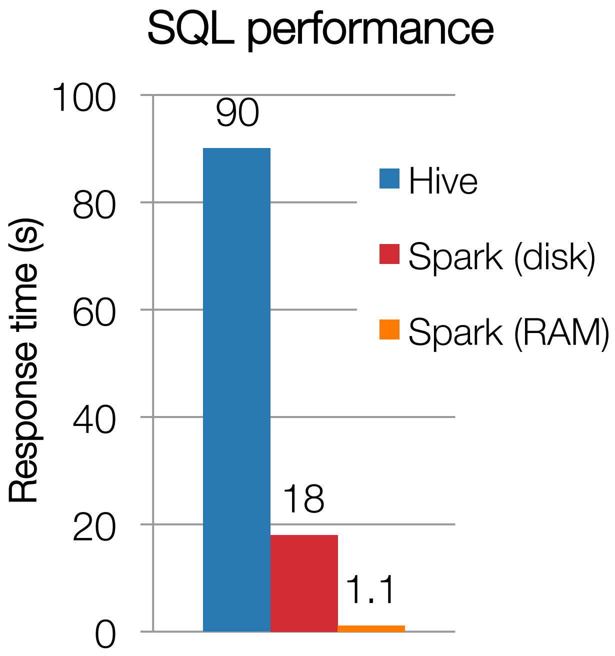
Sophistication: can you run the best algorithms on the data?

These factors have usually required separate, non-commodity tools

Apache Spark

A compute engine for Hadoop data that is:

Fast: up to 100x faster than MapReduce

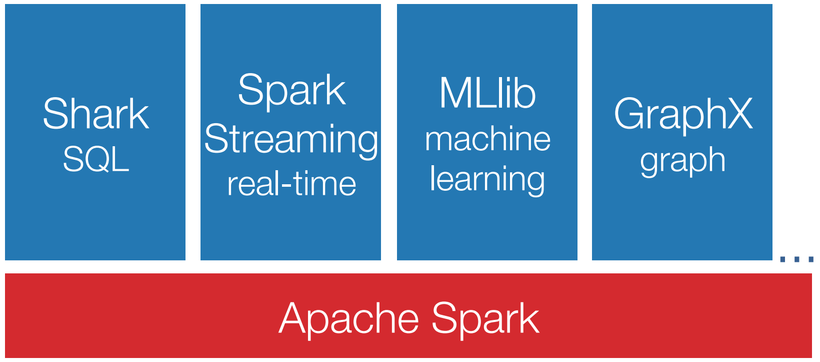


Apache Spark

A compute engine for Hadoop data that is:

Fast: up to 100x faster than MapReduce

Sophisticated: can run today's most advanced algorithms



Apache Spark

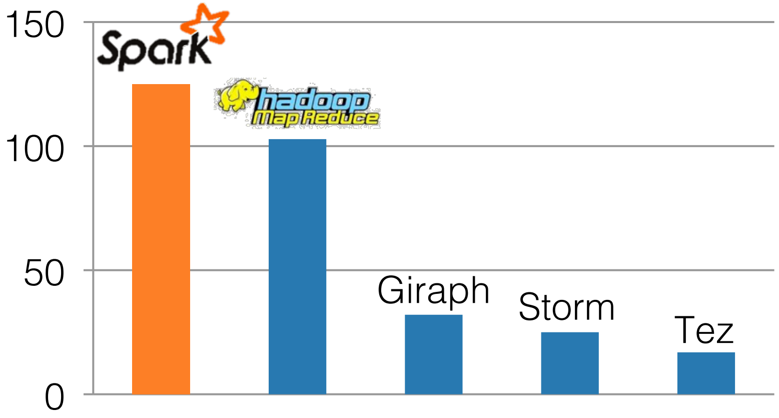
A compute engine for Hadoop data that is:

Fast: up to 100x faster than MapReduce

Sophisticated: can run today's most advanced algorithms

Fully open source: one of most active projects in big data

Contributors in past year



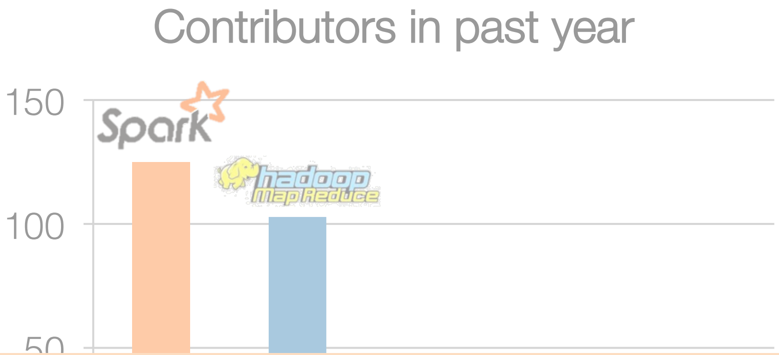
Apache Spark

A compute engine for Hadoop data that is:

Fast: up to 100x faster than MapReduce

Sophisticated: can run today's most advanced algorithms

Fully open source: one of most active projects in big data



Spark brings top-end data analysis to commodity Hadoop clusters

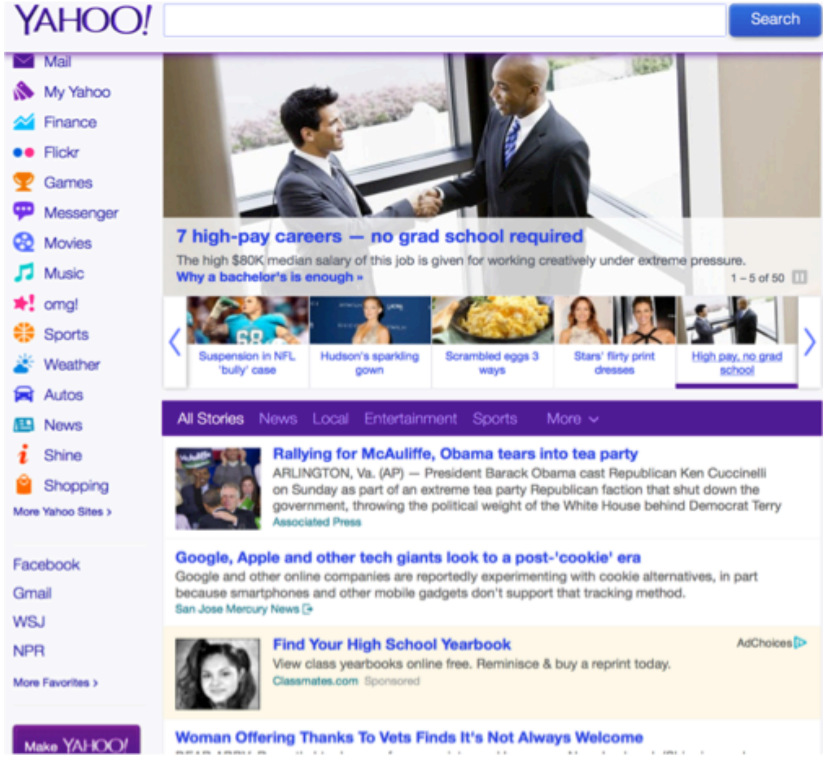
Spark Use Cases

1. Yahoo! Personalization

Yahoo! properties are highly personalized to maximize relevance

Reaction must be **fast**, as stories, etc change in time

Best algorithms are highly **sophisticated**



7 high-pay careers — no grad school required

The high $80K median salary of this job is given for working creatively under extreme pressure.

Why a bachelor's is enough »

Rallying for McAuliffe, Obama tears into tea party

ARLINGTON, Va. (AP) — President Barack Obama cast Republican Ken Cuccinelli on Sunday as part of an extreme tea party Republican faction that shut down the government, throwing the political weight of the White House behind Democrat Terry

Associated Press

Google, Apple and other tech giants look to a post-'cookie' era

Google and other online companies are reportedly experimenting with cookie alternatives, in part because smartphones and other mobile gadgets don't support that tracking method.

San Jose Mercury News

Find Your High School Yearbook

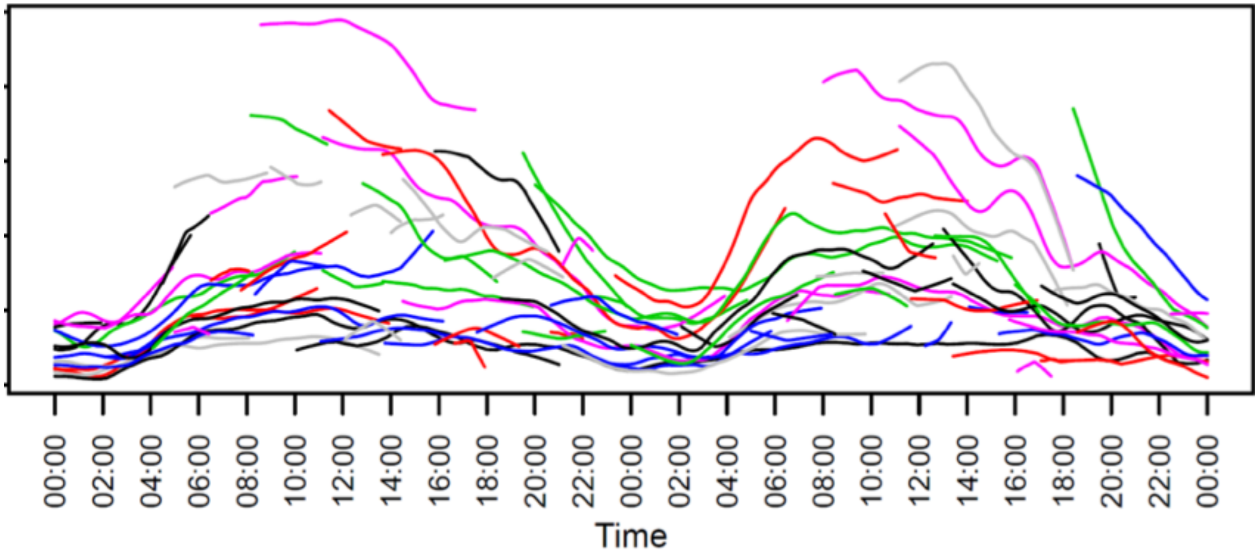
View class yearbooks online free. Reminisce & buy a reprint today.

Classmates.com Sponsored

Woman Offering Thanks To Vets Finds It's Not Always Welcome

1. Yahoo! Personalization

Example challenge: relevance of news stories



Relevance models must be updated throughout the day

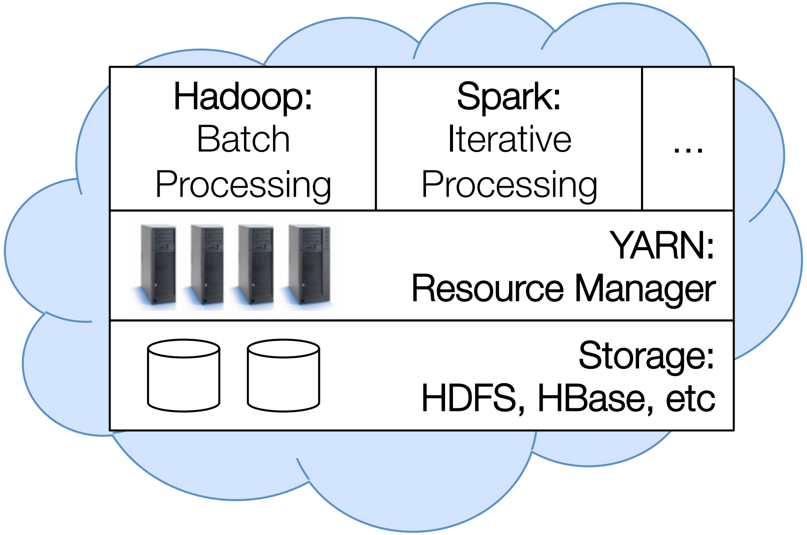
1. Yahoo! Personalization

Spark at Yahoo!

- Runs in Hadoop YARN to use existing data & clusters

Result: pilot for stream ads

- 120 lines in Scala, compared to 15K in C++
- 30 min to run on 100 million samples



Major contributor on YARN support, scalability, operations

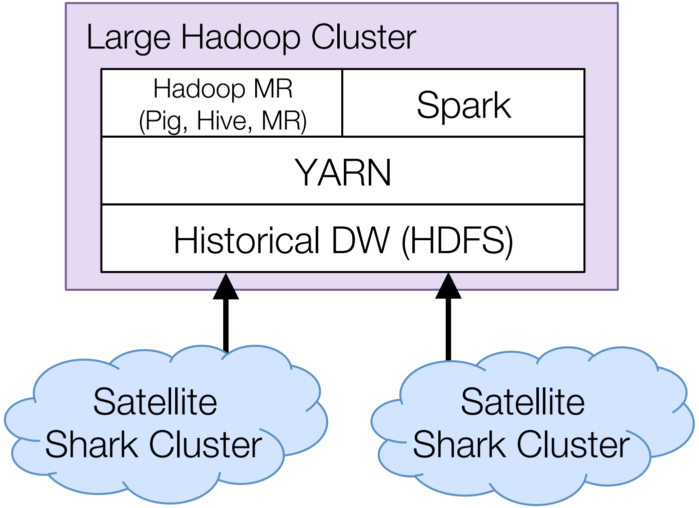
2. Yahoo! Ad Analytics

Yahoo! Ads wanted interactive BI on terabytes of data

Chose Shark (Hive on Spark) to provide this through standard Hive server API + Tableau

Result: interactive-speed queries on terabytes from Tableau

Major contributor on columnar compression, statistics, JDBC

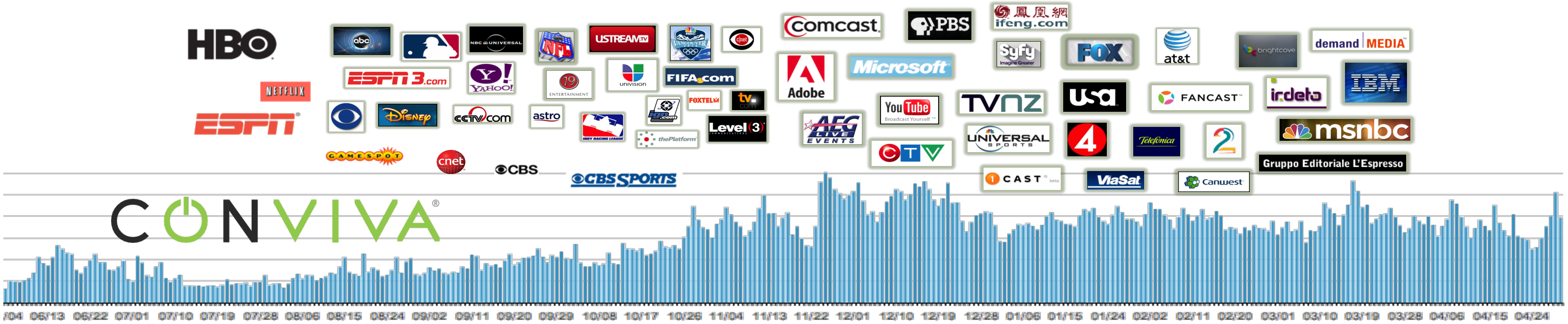


3. Conviva Real-Time Video Optimization

Conviva manages 4+ billion video streams per month

Dynamically selects sources to optimize quality

Time is critical: 1 second buffering = lost viewers

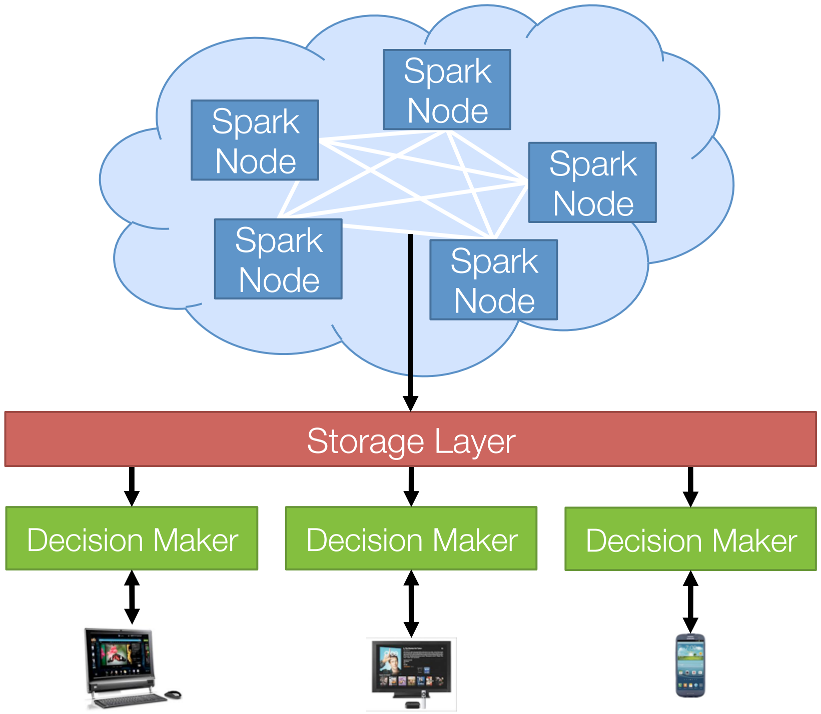


3. Conviva Real-Time Video Optimization

Using **Spark Streaming**, Conviva learns network conditions in real time

Results fed directly to video players to optimize streams

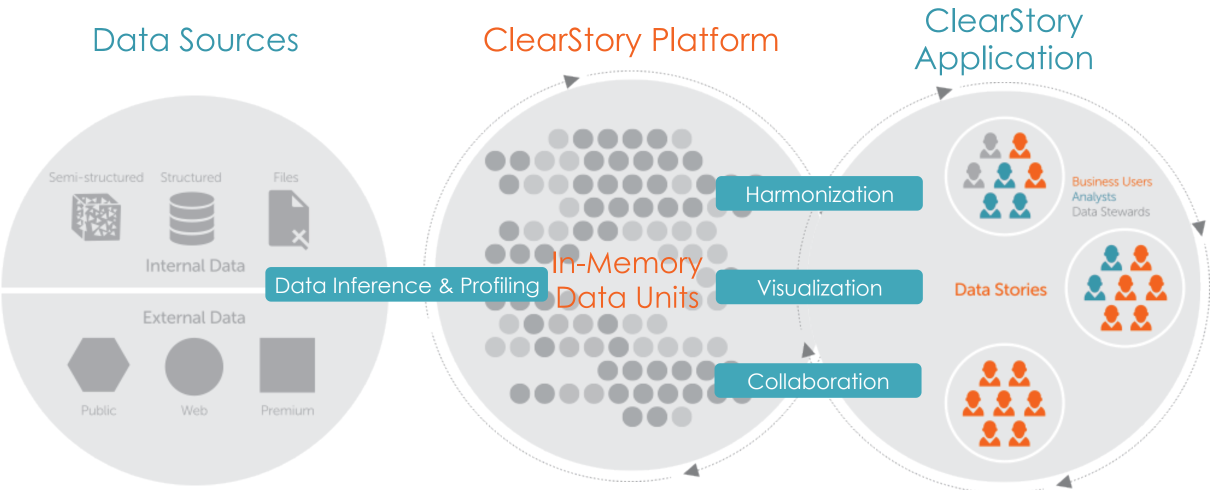
System running in production



4. ClearStory Data: Multi-source, Fast-cycle Analysis

Same-day results from data updating at disparate sources

Dozens of disparate sources converged in seconds/minutes

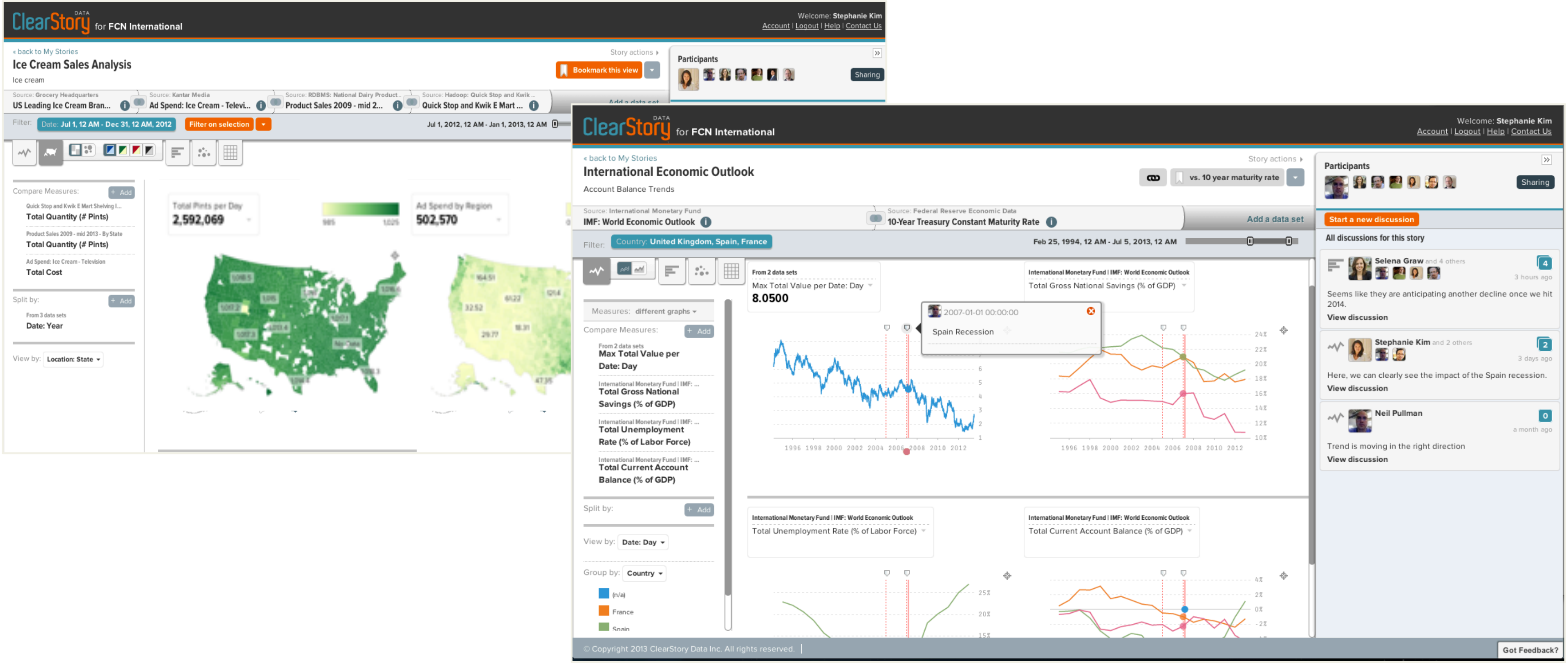


ClearStory DATA

Now You See It™

clearstorydata.com

4. ClearStory Data:
Multi-source, Fast-cycle Analysis



ClearStory DATA for FCN International

Welcome: Stephanie Kim

« back to My Stories

Ice Cream Sales Analysis

Ice cream

Source: Grocery Headquarters — US Leading Ice Cream Bran...

Source: Kantar Media — Ad Spend: Ice Cream - Televi...

Source: RDBMS: National Dairy Product... — Product Sales 2009 - mid 2...

Source: Hadoop: Quick Stop and Kwik ... — Quick Stop and Kwik E Mart ...

Filter: Date: Jul 1, 12 AM - Dec 31, 12 AM, 2012

Jul 1, 2012, 12 AM - Jan 1, 2013, 12 AM

Compare Measures:

- Quick Stop and Kwik E Mart Shelving I... — Total Quantity (# Pints)
- Product Sales 2009 - mid 2013 - By State — Total Quantity (# Pints)
- Ad Spend: Ice Cream - Television — Total Cost

Split by: From 3 data sets — Date: Year

View by: Location: State

Total Pints per Day: 2,592,069

Ad Spend by Region: 502,570

ClearStory DATA for FCN International

Welcome: Stephanie Kim

« back to My Stories

International Economic Outlook

Account Balance Trends

vs. 10 year maturity rate

Source: International Monetary Fund — IMF: World Economic Outlook

Source: Federal Reserve Economic Data — 10-Year Treasury Constant Maturity Rate

Filter: Country: United Kingdom, Spain, France

Feb 25, 1994, 12 AM - Jul 5, 2013, 12 AM

Measures: different graphs

Compare Measures:

- From 2 data sets — Max Total Value per Date: Day
- International Monetary Fund | IMF: ... — Total Gross National Savings (% of GDP)
- International Monetary Fund | IMF: ... — Total Unemployment Rate (% of Labor Force)
- International Monetary Fund | IMF: ... — Total Current Account Balance (% of GDP)

Split by:

View by: Date: Day

Group by: Country

- 🟦 (n/a)
- 🟧 France
- 🟩 Spain

From 2 data sets — Max Total Value per Date: Day — 8.0500

2007-01-01 00:00:00 — Spain Recession

International Monetary Fund | IMF: World Economic Outlook — Total Gross National Savings (% of GDP)

International Monetary Fund | IMF: World Economic Outlook — Total Unemployment Rate (% of Labor Force)

International Monetary Fund | IMF: World Economic Outlook — Total Current Account Balance (% of GDP)

Participants

All discussions for this story

Selena Graw and 4 others — 4 — 3 hours ago

Seems like they are anticipating another decline once we hit 2014.

View discussion

Stephanie Kim and 2 others — 2 — 3 days ago

Here, we can clearly see the impact of the Spain recession.

View discussion

Neil Pullman — 0 — a month ago

Trend is moving in the right direction

View discussion

© Copyright 2013 ClearStory Data Inc. All rights reserved.

Get Started

Download and resources: spark.incubator.apache.org

Free video tutorials: spark-summit.org/2013

Commercial support:

DATABRICKS + cloudera®

Conclusion

Big data will be standard: everyone will have it

Organizations will gain an edge through **speed** of action and **sophistication** of analysis

Apache Spark brings these to Hadoop clusters

DATABRICKS