Big Data: Data Wrangling Boot Camp
R Sentiment Analysis Wrapup

Chuck Cartledge, PhD

24 February 2018

Table of contents (1 of 1)

What are we going to cover?

- Things that we've done
- How to scale the system

How to measure success?

1. Downloaded tweets in real-time
2. Stored tweets for future analysis
3. Conducted sentiment analysis on tweets
4. Displayed sentiment analysis in different ways

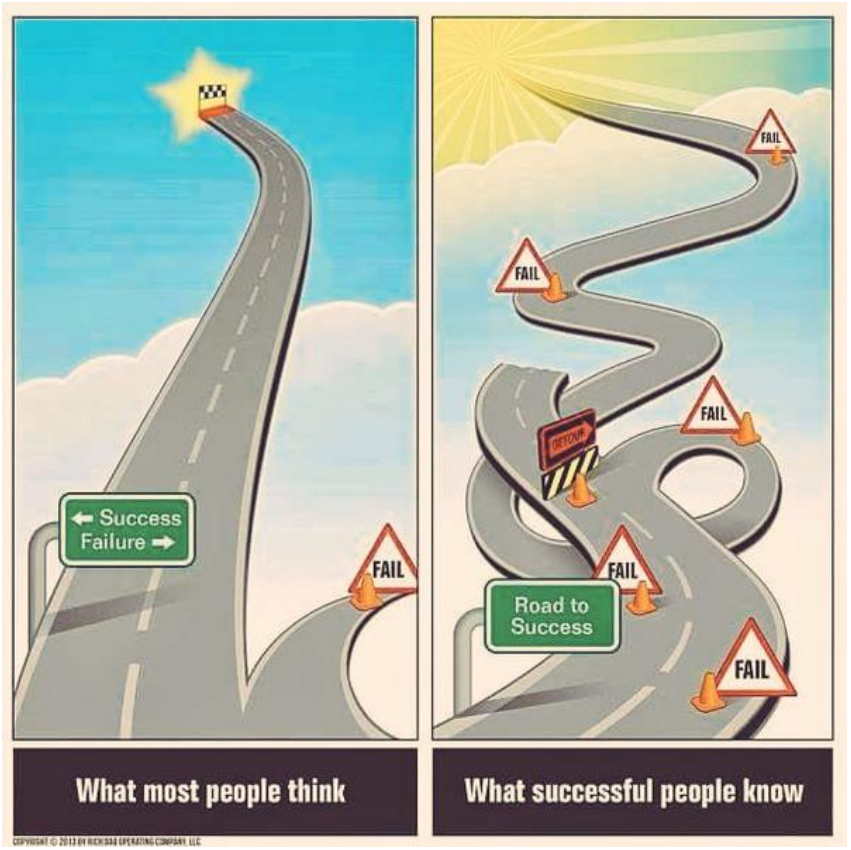
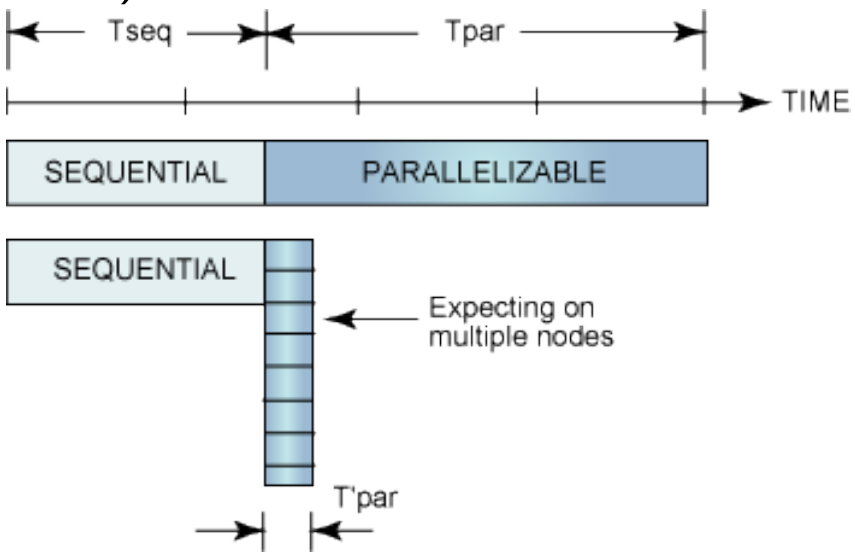


Image from [1].

Amdahl's Law (A summary)

Division and measurement of serial and parallel operations appears time and again. (Shades of Mandelbrot.)

- "Make the common fast."
- "Make the fast common."
- Understand what parts have to be done serially
- Understand what parts can be done in parallel

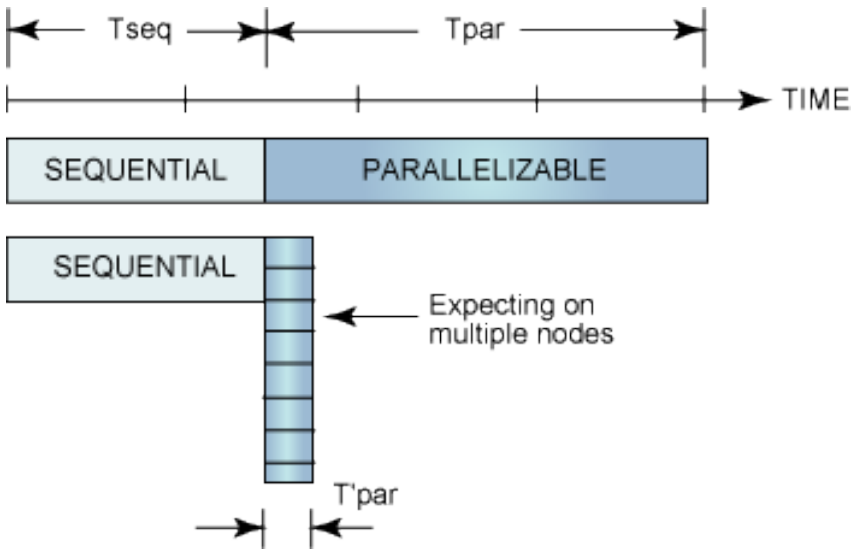


Need to factor in "overhead" costs when computing speed up.

Some details

In our system:

- Parallelizable – data collection, each hashtag could be assigned its own program
- Parallelizable – scheduling of data collection based on hashtag (have the OS do the work)
- Parallelizable – parsing of data based on "known" questions
- Sequential – data storage
- Sequential – data analysis and display



Virtualization as a testing environment

Things to remember

- Virtualization as proof of concept
- Not as fast as real hardware
- Commands and control almost usable

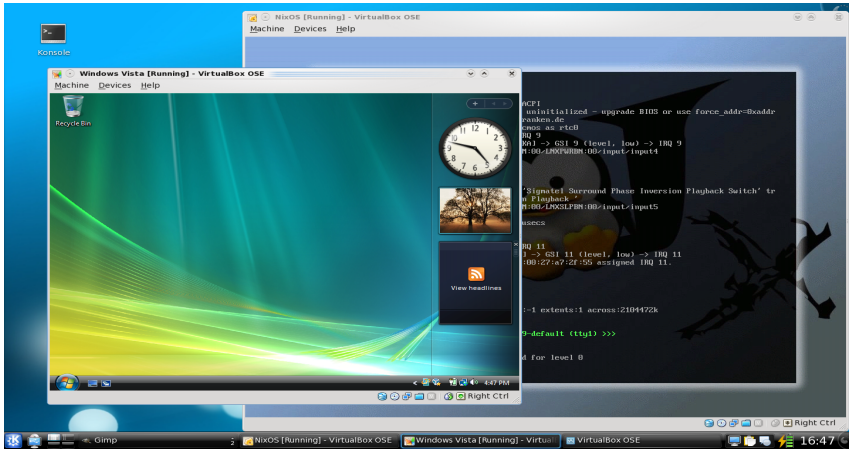


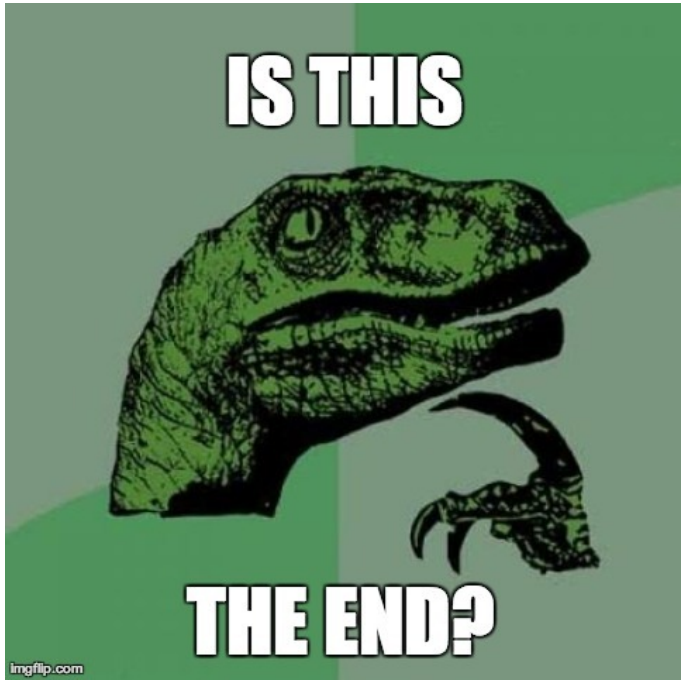
Image from [2].

Q & A time.

Q: What is the burning question on the mind of every dyslexic existentialist?
A: “Is there a dog?”

What have we covered?

- Used our twitter developer account
- **Data wrangled** using R
- Conducted sentiment analysis on live tweets
- Looked at the sentiment analysis in different ways



Next: Scraping static web pages.

References (1 of 1)

[1] Sina H, Road to success, https://thedailyteacher.com/2016/05/20/the-road-to-success/, 2016.

[2] NixOS Staff, Nixos screenshots, https://nixos.org/nixos/screenshots.html, 2016.