Big Data: Data Wrangling Boot Camp
Python Sentiment Analysis Wrapup

Chuck Cartledge, PhD

17 September 2016

Table of contents I

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

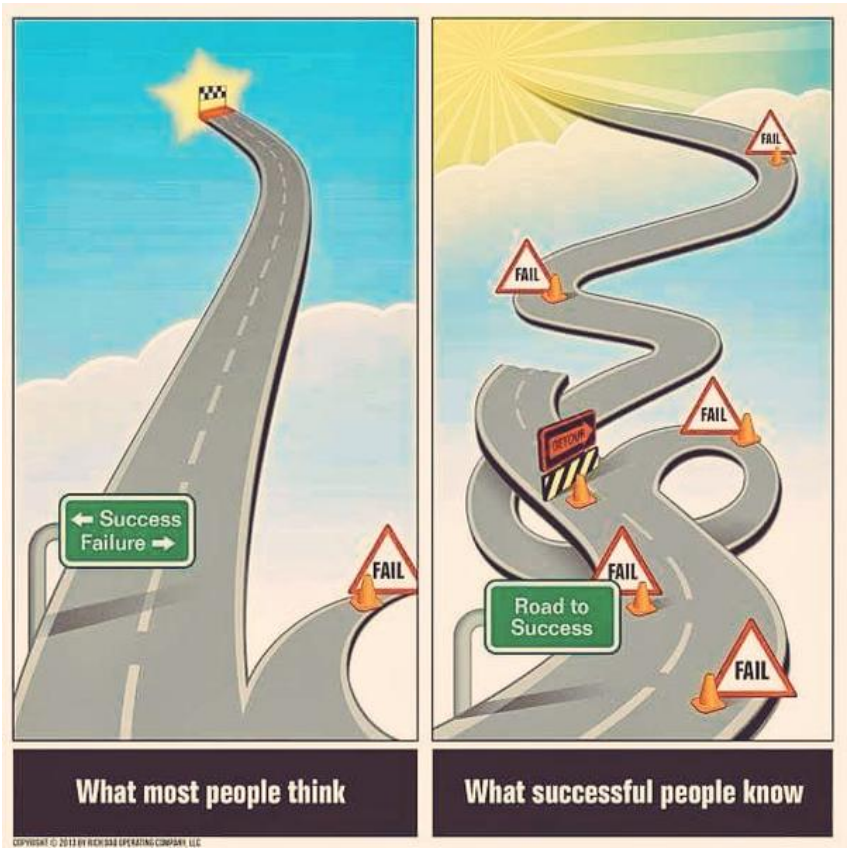
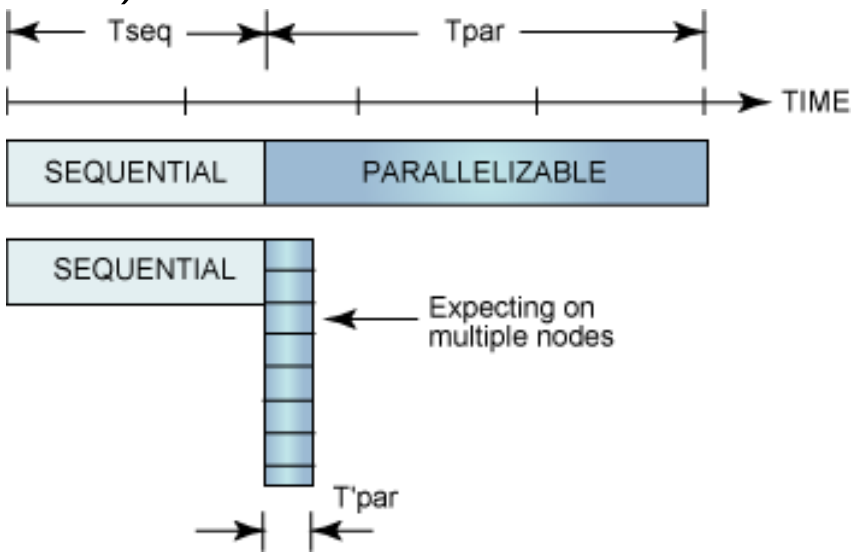


Image from [1].

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

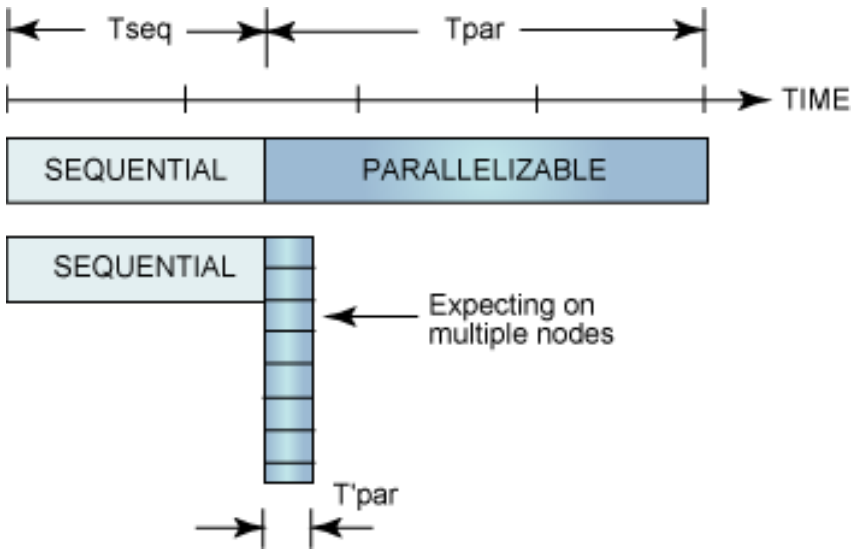


Need to factor in "overhead" costs when computing speed up.

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

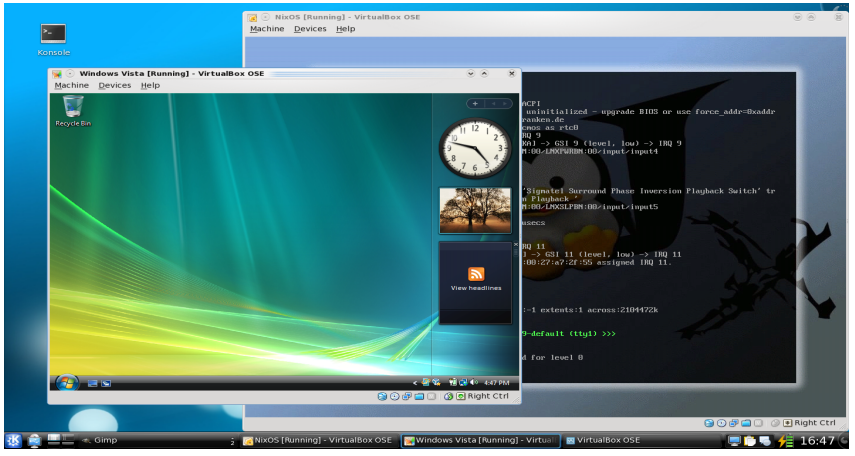


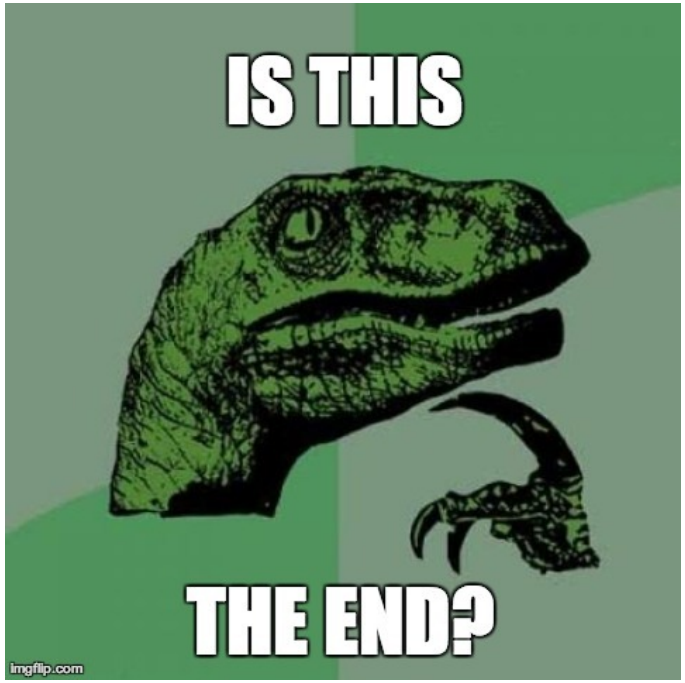
Image from [2].

Q & A time.

"'The Answer to the Great Question . . . Of Life, the Universe and Everything . . . is . . . forty-two,' said Deep Thought, with infinite majesty and calm."
Douglas Adams, The Hitchhiker's Guide to the Galaxy

What have we covered?

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



Next: Hands on analysing tweets with R.

References I

[1] jhoncarter071, Road to success, http://imgur.com/gallery/nBnkz06, 2016.

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