Cloud-Based Text Analytics: Harvesting, Cleaning and Analyzing Corporate Earnings Conference Calls

Michael Chuancai Zhang[1], Vikram Gazula[2], Dan Stone[1], Hong Xie[1]

1 Gatton College of Business and Economics, 2 Center for Computational Sciences, University of Kentucky

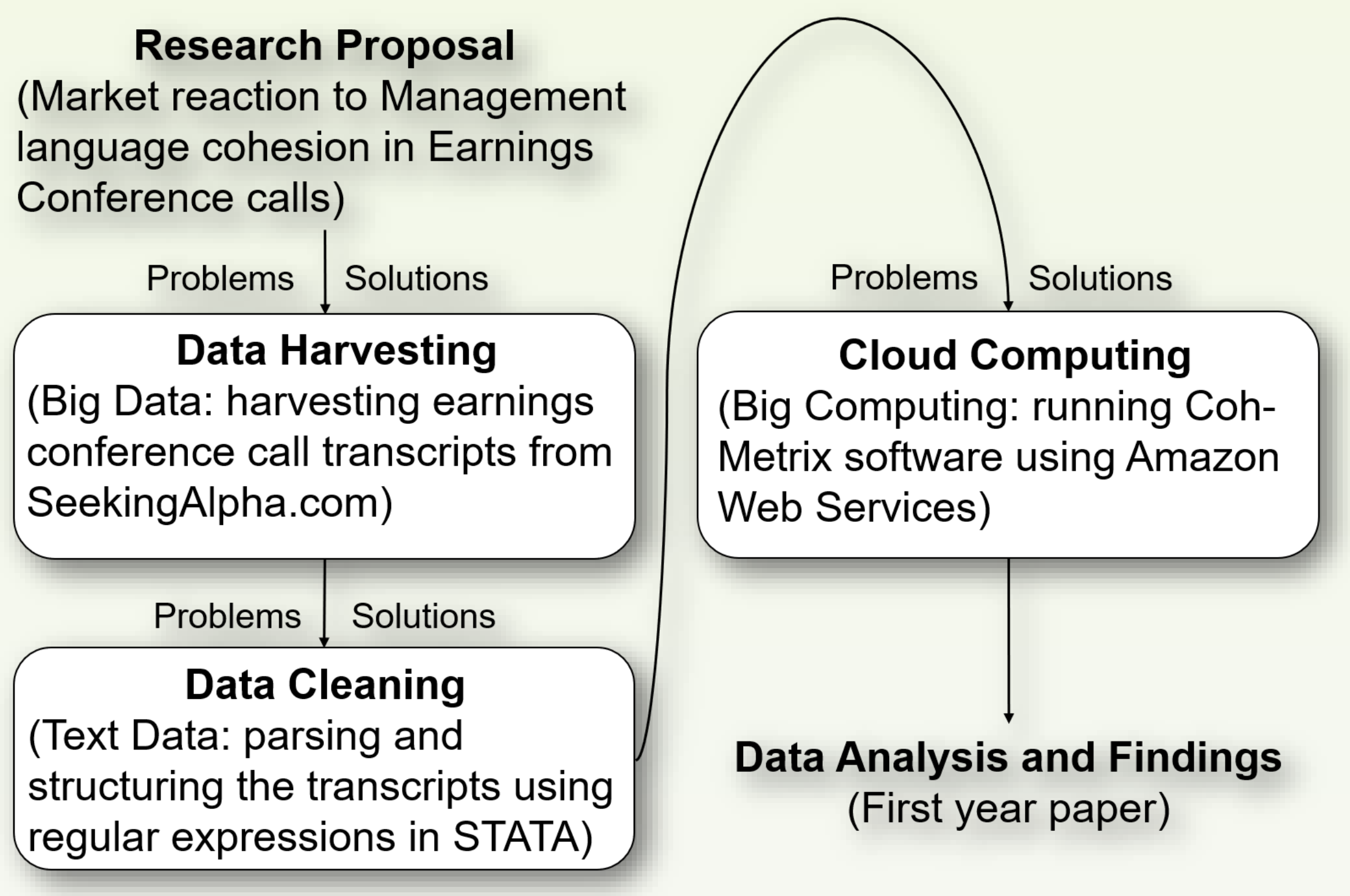
Background

Does management language cohesion in earnings conference calls matter to the capital market? As a part of the research on the above question, and taking advantage of the modern IT technologies, this project:

- harvested 115,882 earnings conference call transcripts from SeekingAlpha.com
- parsed and structured 89,988 transcripts using regular expressions in Stata
- analyzed 179,976 text files using Amazon Elastic Compute Cloud (Amazon EC2), which
- saved almost 2 years (675 days) of the project time

As this project is related to big data, text analytics, and big computing, it may be a good case to show how we can benefit from modern computation technologies in our research.

Framework



Research Proposal
(Market reaction to Management language cohesion in Earnings Conference calls)

Problems | Solutions

Data Harvesting
(Big Data: harvesting earnings conference call transcripts from SeekingAlpha.com)

Problems | Solutions

Data Cleaning
(Text Data: parsing and structuring the transcripts using regular expressions in STATA)

Problems | Solutions

Cloud Computing
(Big Computing: running Coh-Metrix software using Amazon Web Services)

Data Analysis and Findings
(First year paper)

Figure 1 The framework of this project

Objectives

The objectives of this project are to:

- introduce the problems encountered in the data harvesting, cleaning, and analyzing processes and the final solutions using modern technologies
- provide insights for scholars interested in or planning to do "big text data" related research in other domains
- provide basic knowledge of future technologies (e.g. cloud computing)

Problems and Solutions

♠ Call transcripts harvesting (big data)

- **Jobs**: downloading 30 (transcript links per page) * 4,279 (pages) transcripts (by 4/1/2017) from Seekingalpha.com
- **Problems**: estimated 974 hours (41days) to harvest the transcripts manually
- **Solutions**: automatically and friendly harvesting with a web crawler coded using Stata program (see diagram below)

♠ Call transcripts cleaning (text analytics)

- **Jobs**: parsing the unstructured html transcripts to structured and clean text files without html tags and special characters
- **Problems**: (1) estimated 9,739 hours (406 days) to clean the transcripts manually; (2) dealing with irregular transcripts without full information
- **Solutions**: four stage parsing strategy using regular expressions in Stata (see diagram below)

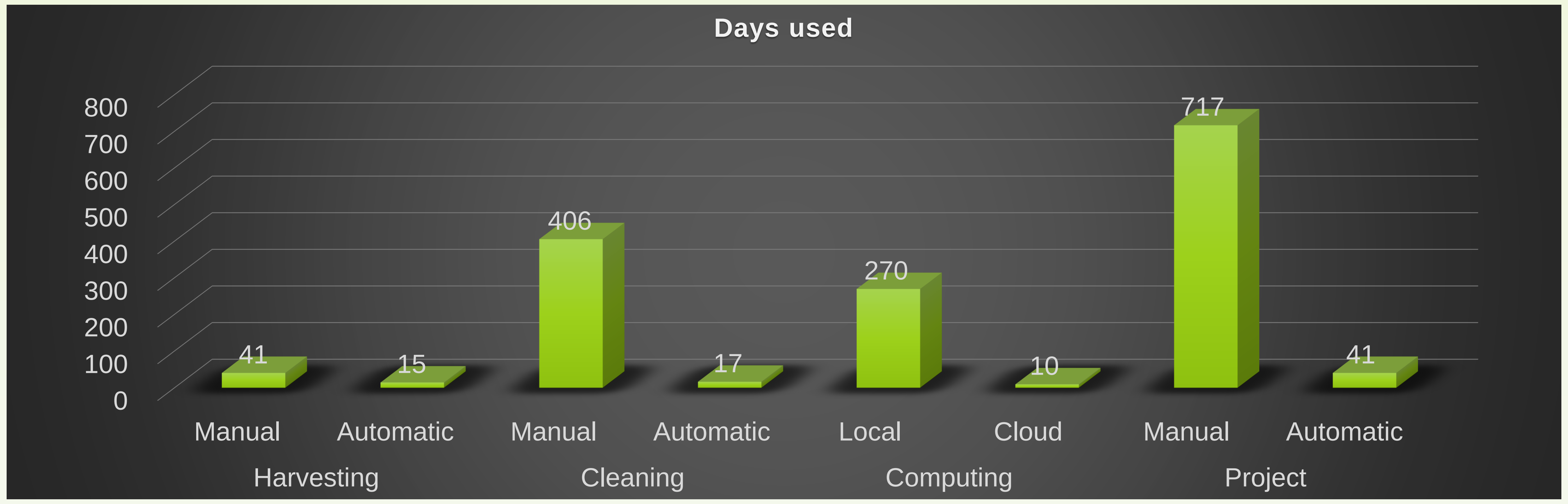


Figure 2 Days used for harvesting, cleaning, and computing the transcripts

Notes: (1) The estimated manually downloading speed is 30 seconds per transcript, including opening the link, selecting and copying the content, opening a text file, past the content, and name and save the file. The estimated automatically downloading speed is 10 seconds per transcript, including 3 minutes waiting time per 20 transcripts because SeekingAlpha.com does not allow continuous downloading (anti crawling techniques).

(2) The estimated manually cleaning speed is 300 seconds per transcript. The actual automatically cleaning speed is 2 seconds per transcripts and the total days include 14 coding days.

(3) The estimated running speed of Coh-Metrix is 8.1 Kilobyte per minute on a local benchmark machine, Lenovo ThinkPad X1 Yoga (i7-6600U CPU; 16G RAM; 64-bite operating system). For the cloud computing using Amazon EC2 from Amazon web services, we choose the c4.8xlarge instances with 36 vCPUs and actually use an average of 27 CPUs in each instance, the optimal solution for our project considering the computing cost.

♠ Call transcripts analyzing (big computing)

- **Jobs**: running Coh-Metrix software (developed by University of Memphis) to get language cohesion measures
- **Problems**: (1) estimated 6,473 hours (270 days) to finish the job using a local machine; (2) cost of buying new machines
- **Solutions**: using Amazon Elastic Compute Cloud (Amazon EC2) from Amazon web services (AWS) (see diagram above)

♠ Cost of AWS EC2 cloud computing

Table 1 Cost of AWS EC2

	On-Demand Instances	Spot Instances
Total files	25,000	130,225
Hours used	82	469
Cost per hour	$3.09	$1.58
Total Cost	$249	$743

Notes: AWS EC2 provides two kinds of instances. On-Demand instances are purchased at a fixed rate per hour and are well-suited for short-term, irregular workloads that cannot be interrupted. Spot Instance prices are typically more than 75% lower than On Demand prices, but will be cut off when the Spot price increases and exceeds the customer's bid. The Spot instances are well-suited for high-performance computing that tolerates interruption.

Achievements

- **Coding solution:** Stata coding solution for downloading and parsing the earnings conference call transcripts
- **Language cohesion data**: 108 Coh-Metrix measures for each earnings conference call transcripts
- **First-year paper**: management language cohesion provides incremental information to capital market
- **Future technology for big computing**: Cloud computing from Amazon Web Services

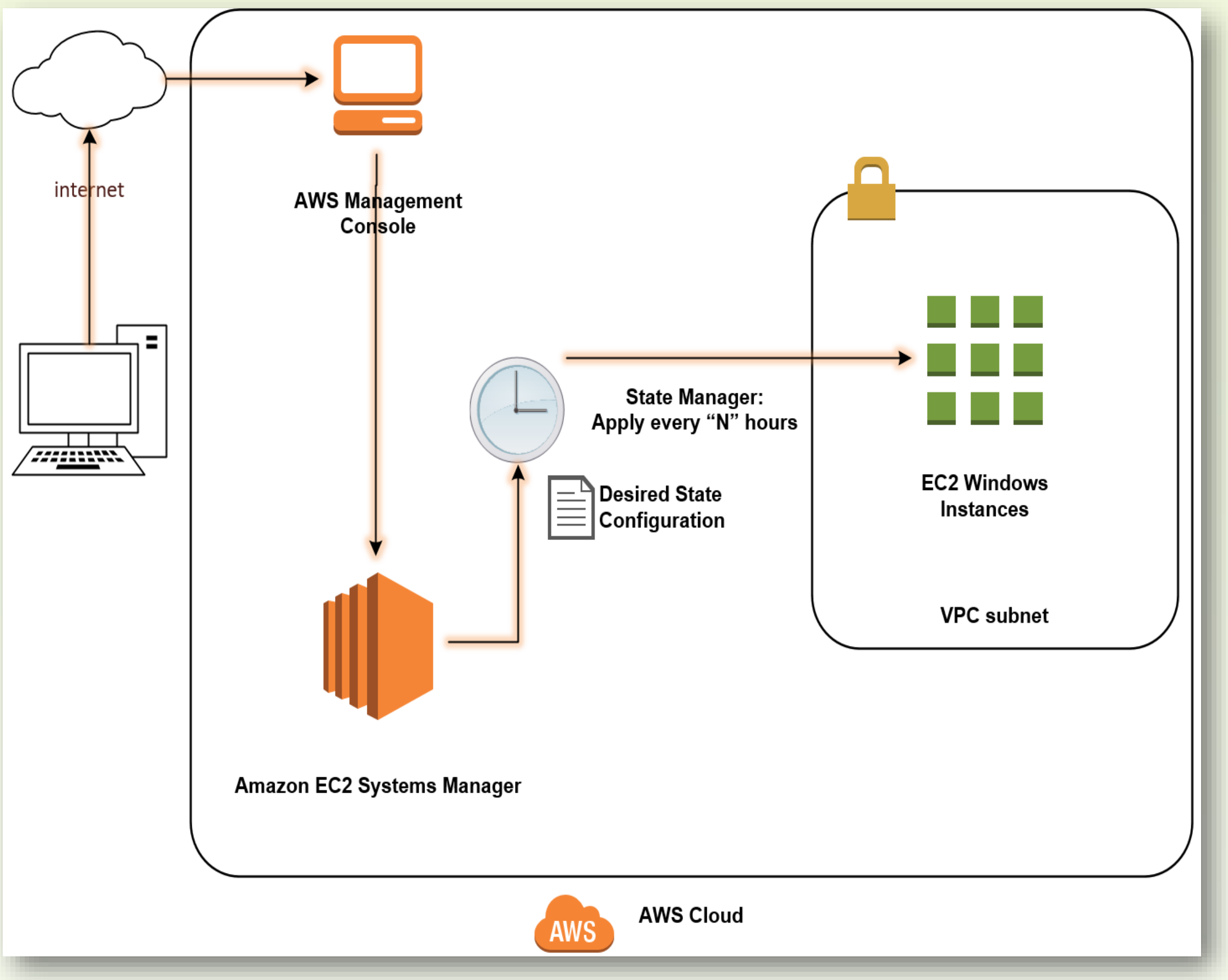

internet
AWS Management Console
State Manager: Apply every "N" hours
Desired State Configuration
EC2 Windows Instances
VPC subnet
Amazon EC2 Systems Manager
AWS
AWS Cloud

Figure 3 An example flow chart of AWS EC2

The figure is from Shaun Breen's blog in AWS Management Tools Blog.

Future work

AWS cloud computing:

- CEO and CFO Language cohesion in earnings conference calls
- Management language cohesion in SEC filings
- Traditional linguistic complexity measures (e.g. Bog index) of earnings conference calls

Acknowledgement

- We thank the Coh-Metrix development team from University of Memphis for providing us the software.
- We thank the AWS team for their IT support
- We gratefully acknowledge the financial support from the University of Kentucky, the Gatton College of Business, and, the Von Allmen School of Accountancy.