Cloud Architectures

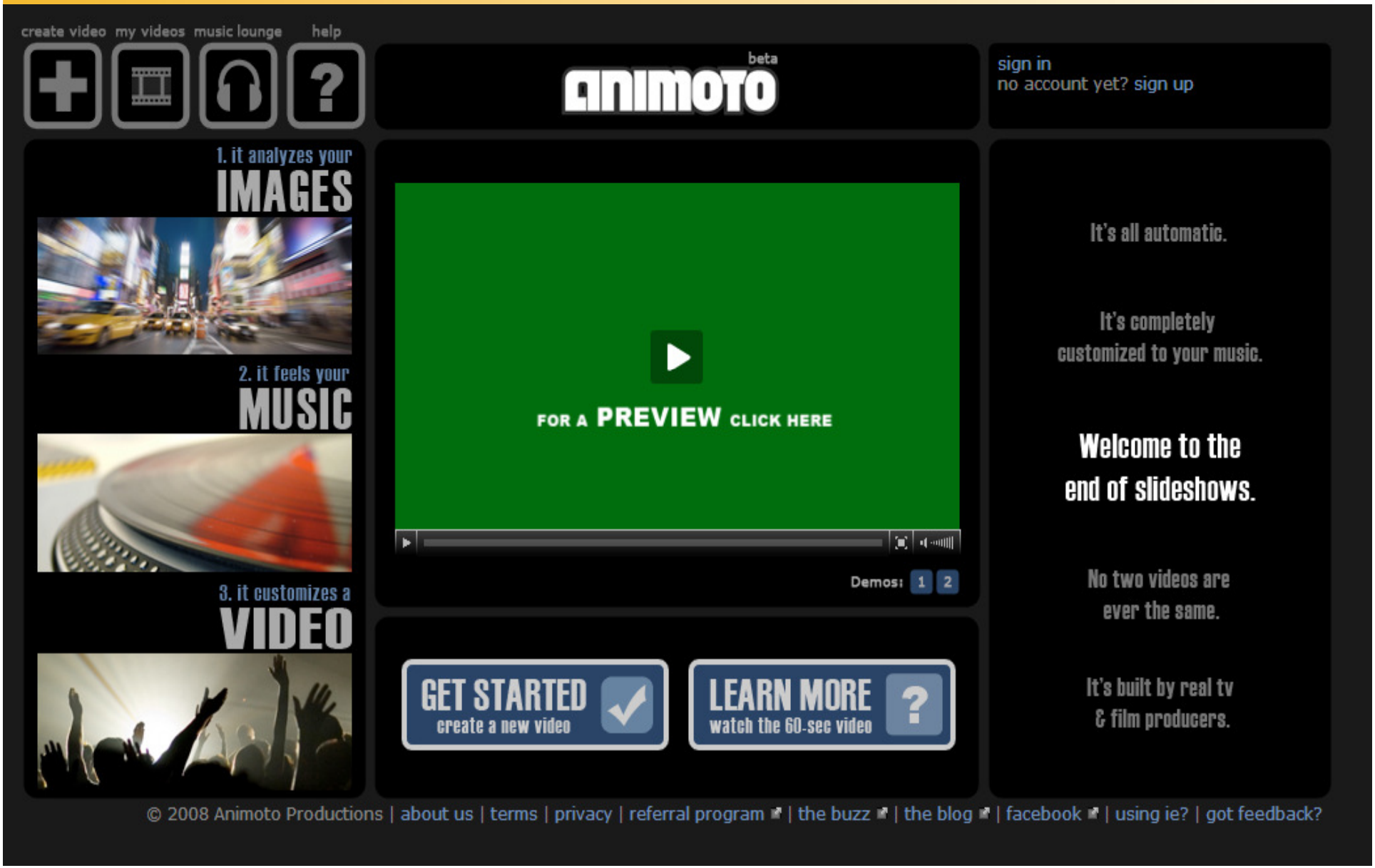
Jinesh Varia

Technology Evangelist
Amazon Web Services

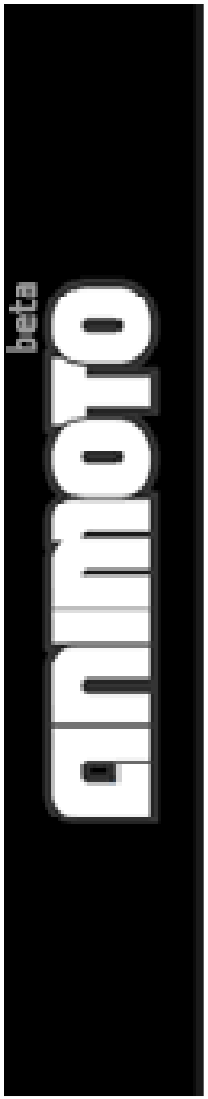
amazon web services™

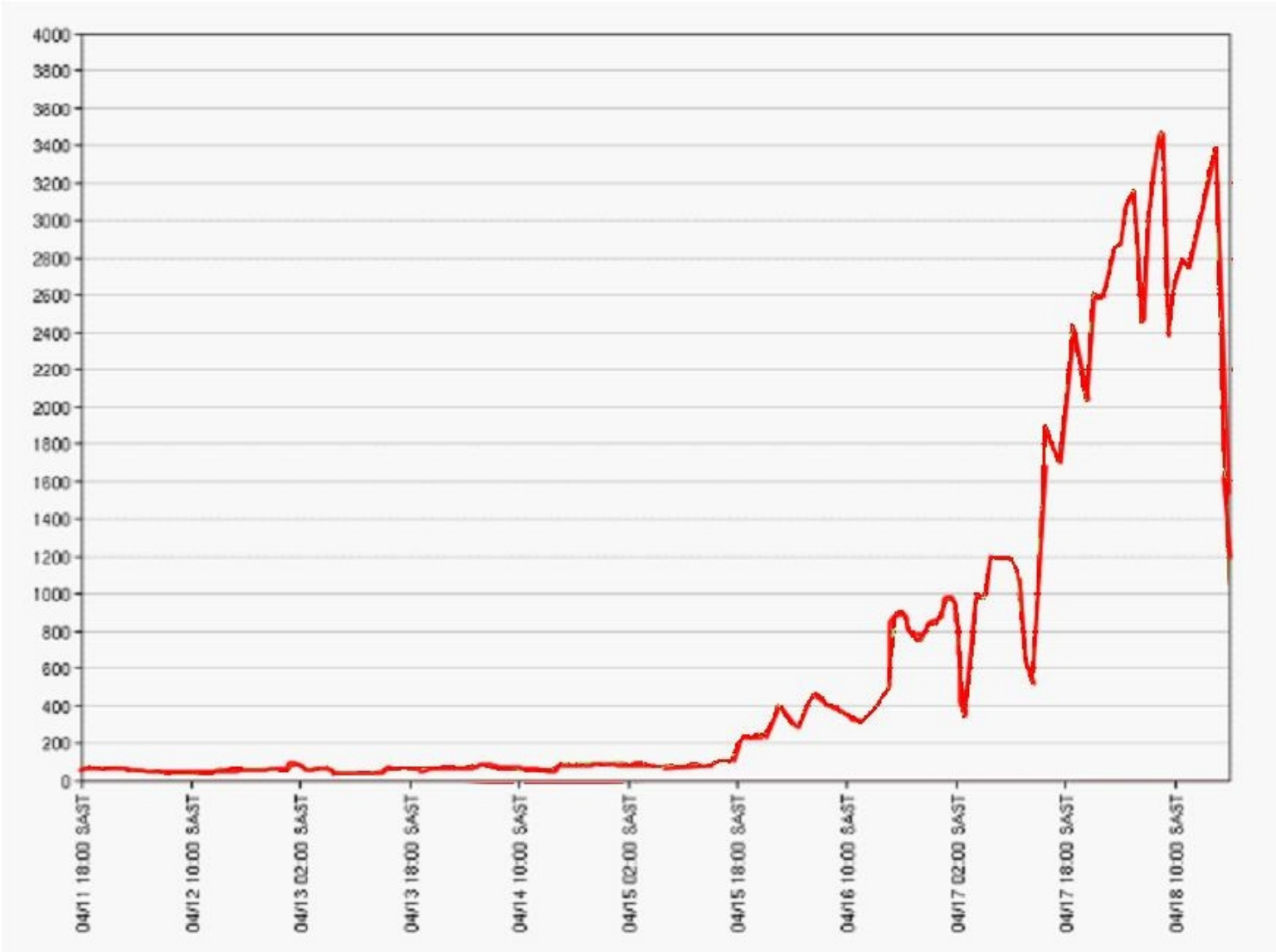
ANIMOTO.COM

amazon web services



Scale: 50 servers to 3500 servers in 3 days





Coding in the Cloud

Culture has changed:
Server-less Startup Companies

amazon web services

amazon web services™

Start-Up Challenge

Winner December 2007

o·o·y·a·l·a™

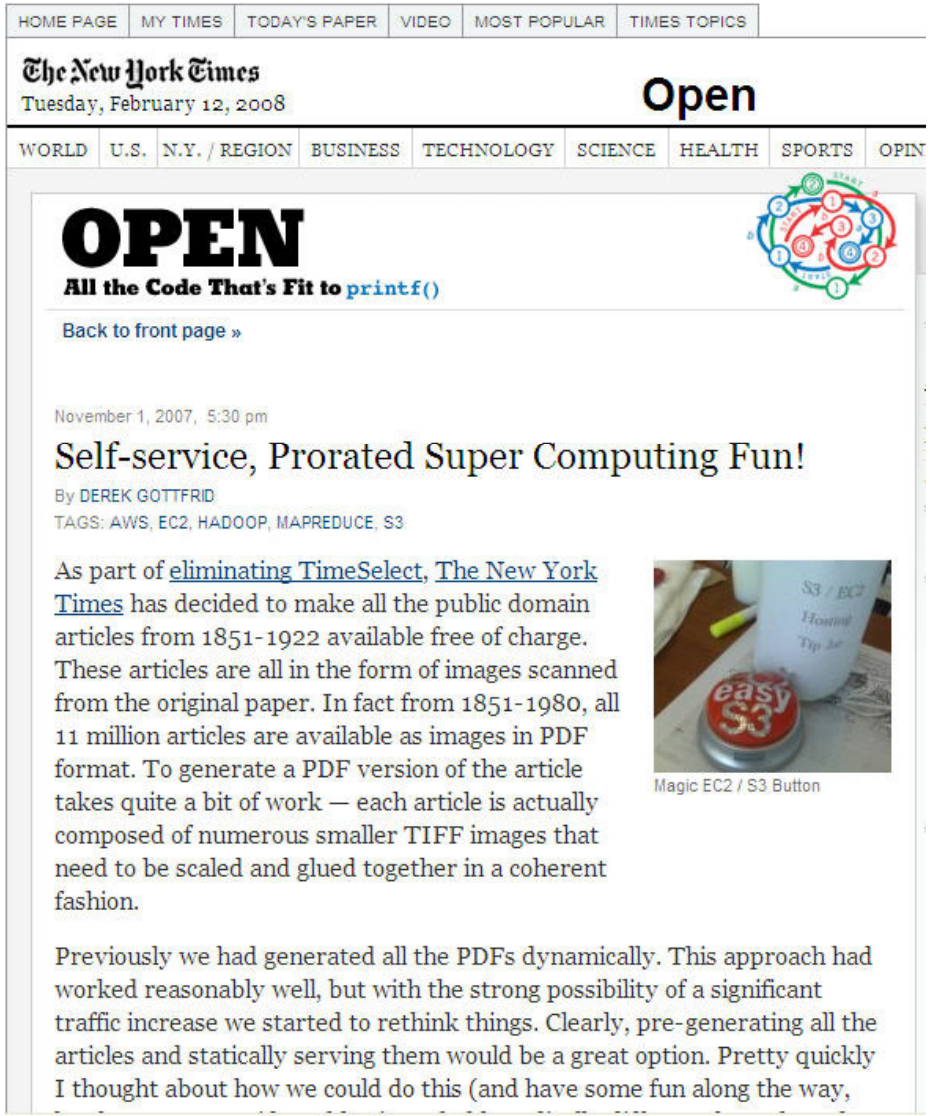
Prize: Golden Hammer

Photo: Smashing the hardware


Culture of Computing has changed


Culture of Computing has changed
Economics of Computing has changed

“TimesMachine” from NY Times



HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times
Tuesday, February 12, 2008

Open

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPIN

OPEN

All the Code That's Fit to printf()

Back to front page »

November 1, 2007, 5:30 pm

Self-service, Prorated Super Computing Fun!

By DEREK GOTTFRID
TAGS: AWS, EC2, HADOOP, MAPREDUCE, S3

As part of eliminating TimeSelect, The New York Times has decided to make all the public domain articles from 1851-1922 available free of charge. These articles are all in the form of images scanned from the original paper. In fact from 1851-1980, all 11 million articles are available as images in PDF format. To generate a PDF version of the article takes quite a bit of work — each article is actually composed of numerous smaller TIFF images that need to be scaled and glued together in a coherent fashion.

Magic EC2 / S3 Button

Previously we had generated all the PDFs dynamically. This approach had worked reasonably well, but with the strong possibility of a significant traffic increase we started to rethink things. Clearly, pre-generating all the articles and statically serving them would be a great option. Pretty quickly I thought about how we could do this (and have some fun along the way,

- 1851-1922 Articles
- TIFF -> PDF
- Input: 11 Million Articles (4TB of data)
- What did he do ?
  - 100 EC2 Instances for 24 hours
  - All data on S3
  - Output: 1.5 TB of Data
  - Hadoop, iText, JetS3t
  - Under $400


USE ONLY
WHAT YOU
NEED.
DENVER WATER


Culture of Computing has changed
Economics of Computing has changed


Culture of Computing has changed
Economics of Computing has changed
Education in Computing has changed

CS290F : Scalable Internet Services

USCB Fall 2006

- Prof created an app to manage team usage
- Ruby on Rails
- Complete Stack: From Load balancer, App Server to DB
- Learn how to scale: Simulated load
- Generated Graphs
- All course contents, students assignments, lessons learned are on the Wiki

CS345a : Data Mining @ Stanford

Tools used:

- Shell/Linux/Java
- Hadoop on EC2
- Data set on S3
- Datasets :NetFlix, Alexa, IR datasets from TREC

Class organization:

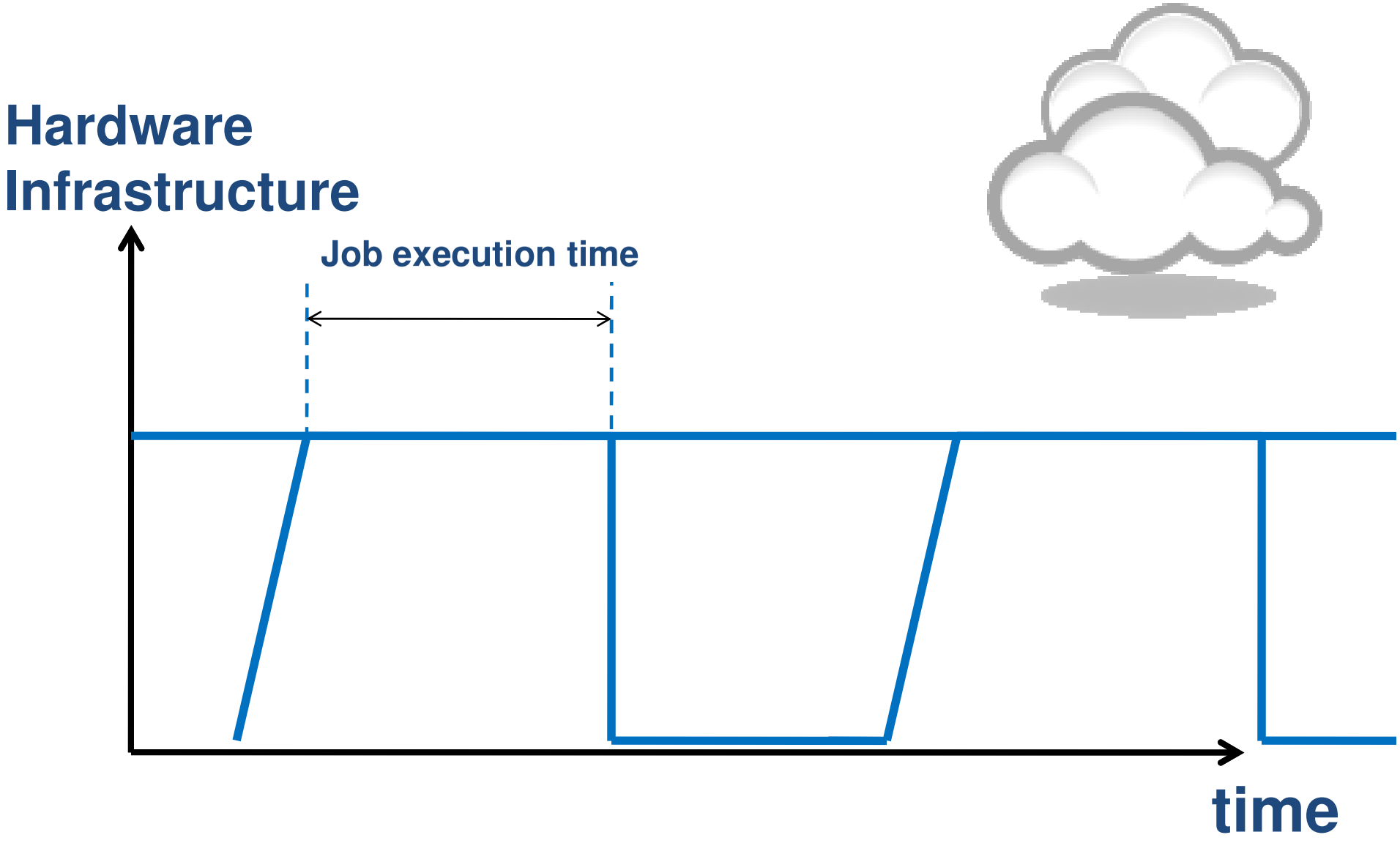
Stanford Winter 2007

- 30-35 Students
- Each Team spawns 10-15 Hadoop slave nodes
- TA created Getting-Started AMIs (& scripts)
- TA managed the students usage

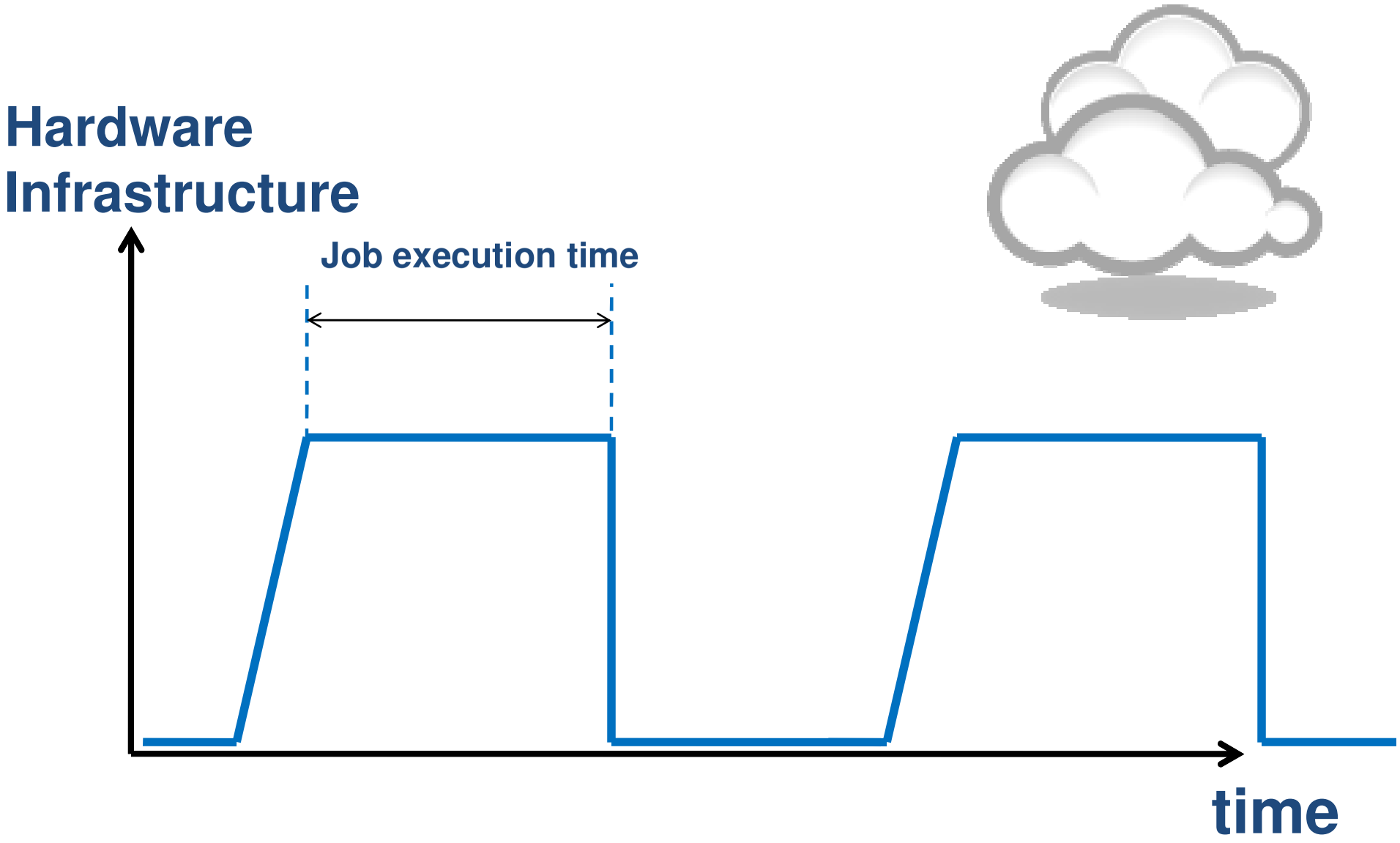

Culture of Computing has changed
Economics of Computing has changed
Education in Computing has changed
Concepts in Computing has changed

Cloud Architectures

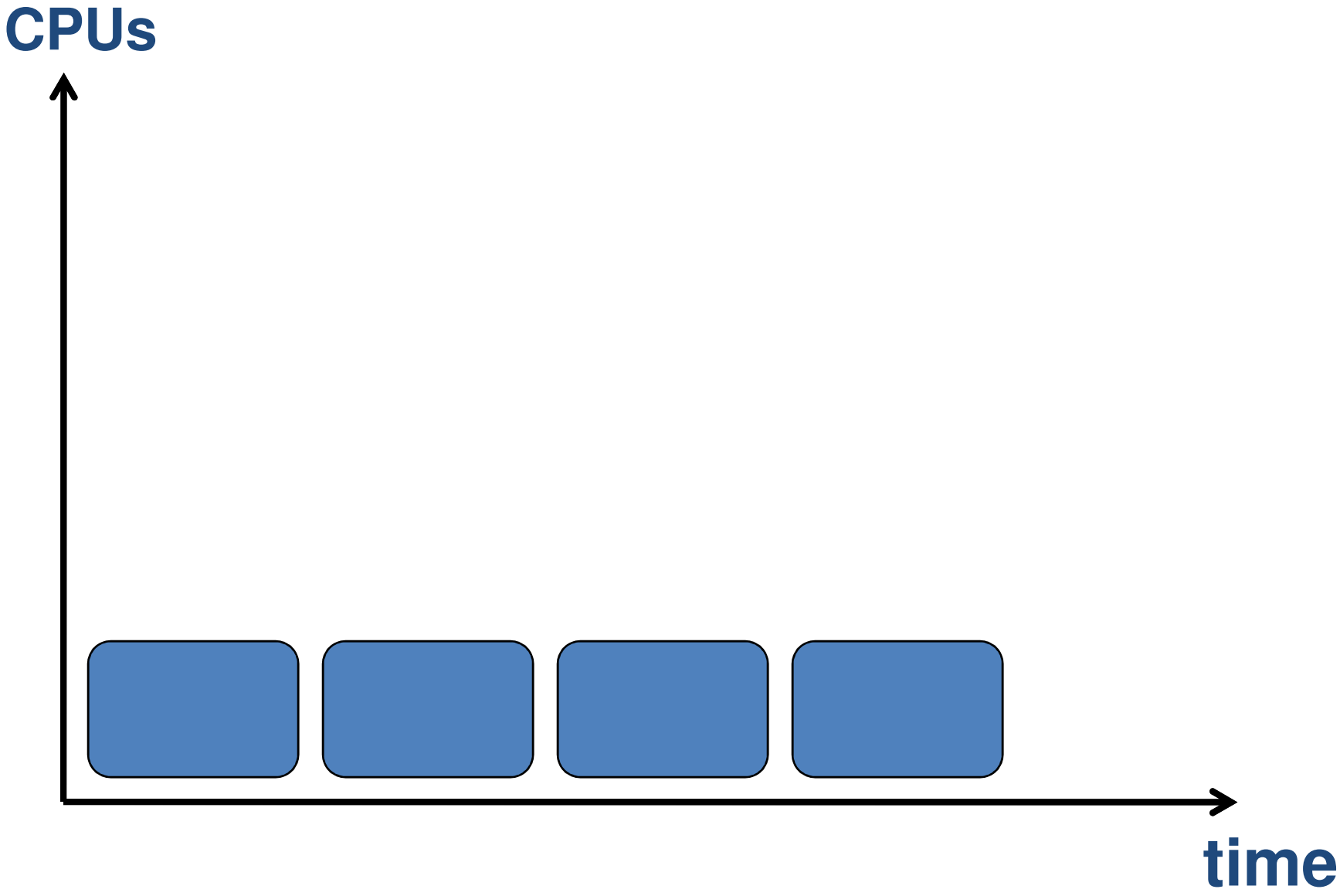
Cloud Architectures



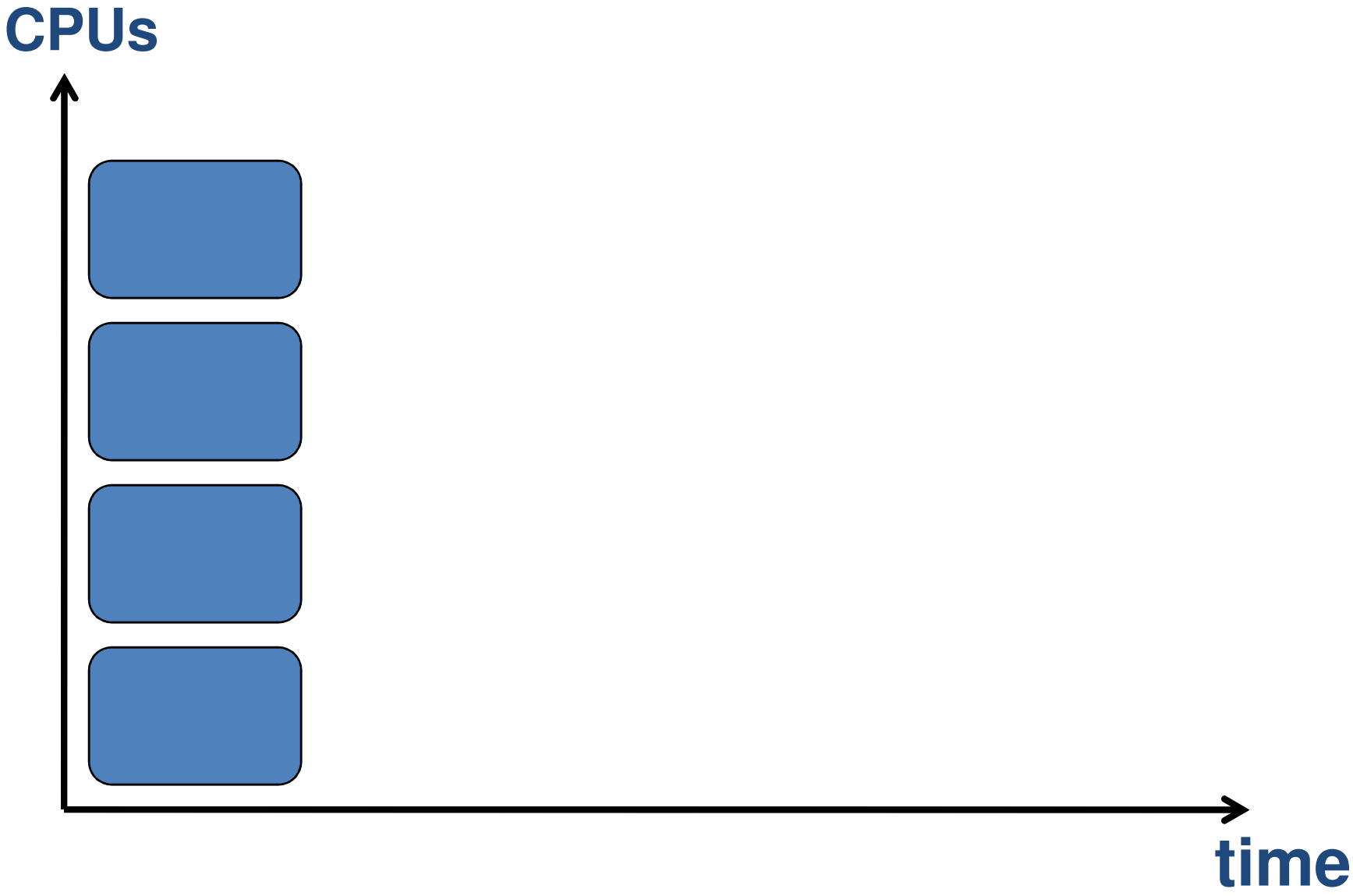
Cloud Architectures



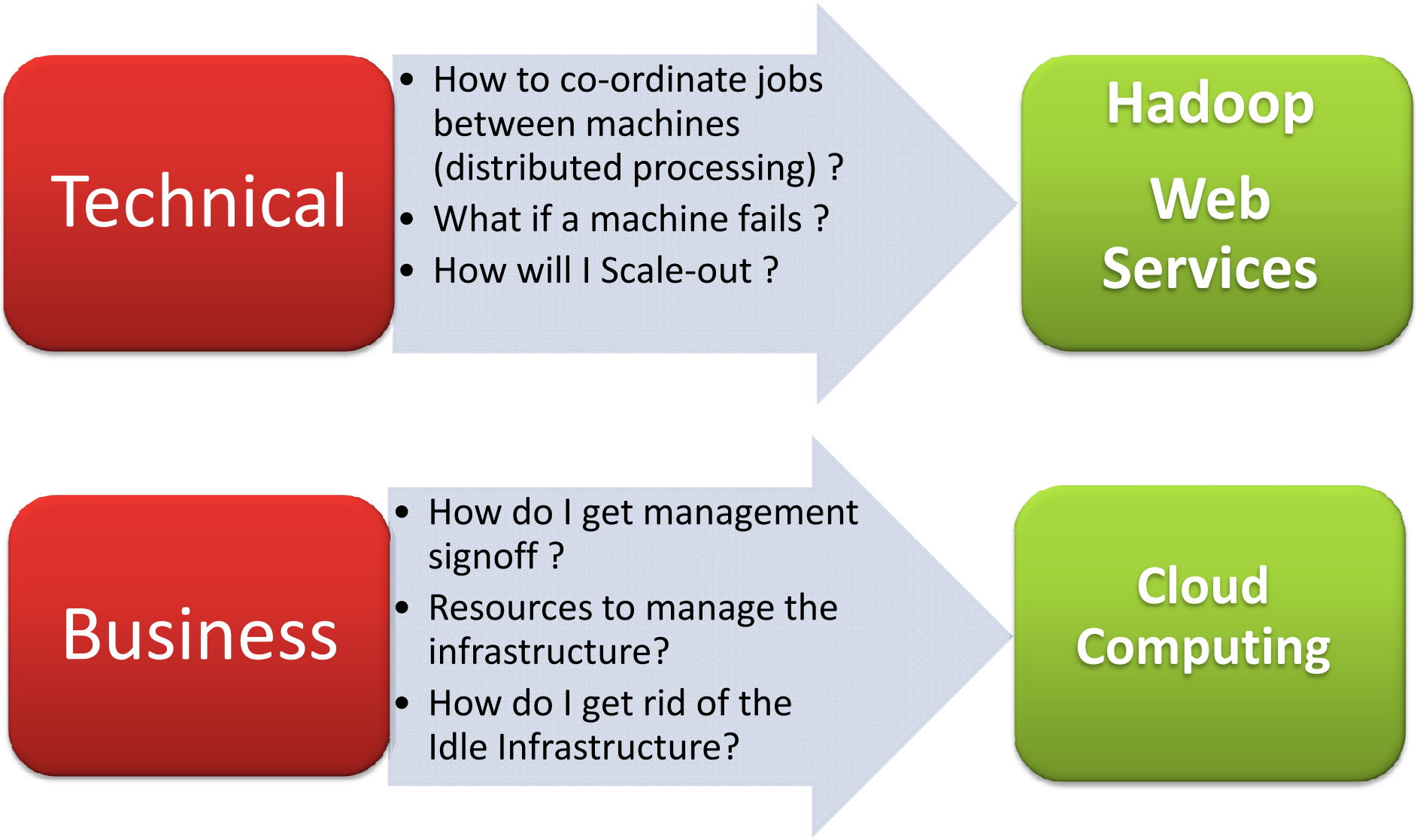
Shrink your processing time



Shrink your processing time



Main Problems



Let's take a usecase…

- *Web Company* : Analyze large-data sets of clickstream logs
- *Social Networking Company* : Analyze demographic and market data
- *Phone Company* : Locate all customers who have called in a given area
- *Large Retailer Chain* : Wants to know what items a particular customer bought last month
- *Surveillance Company* : Wants to transcode video for last several years
- *Pharma Company* : Wants locate people who were prescribed a certain drug

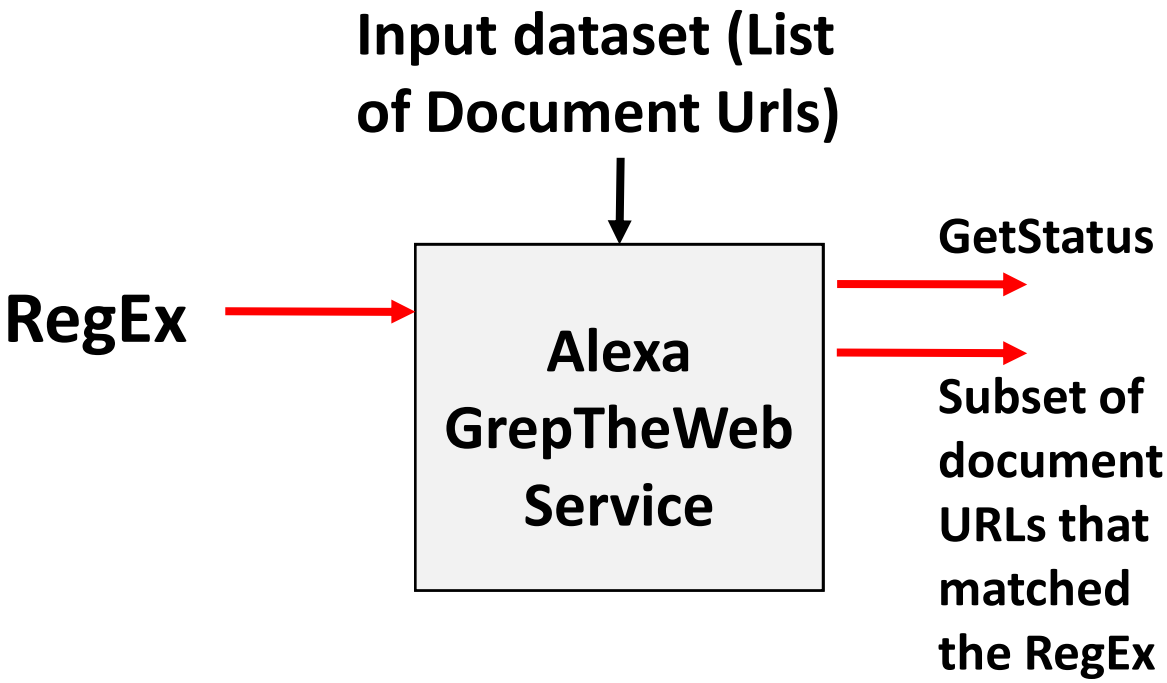
GrepTheWeb

What's so cool about GrepTheWeb ?

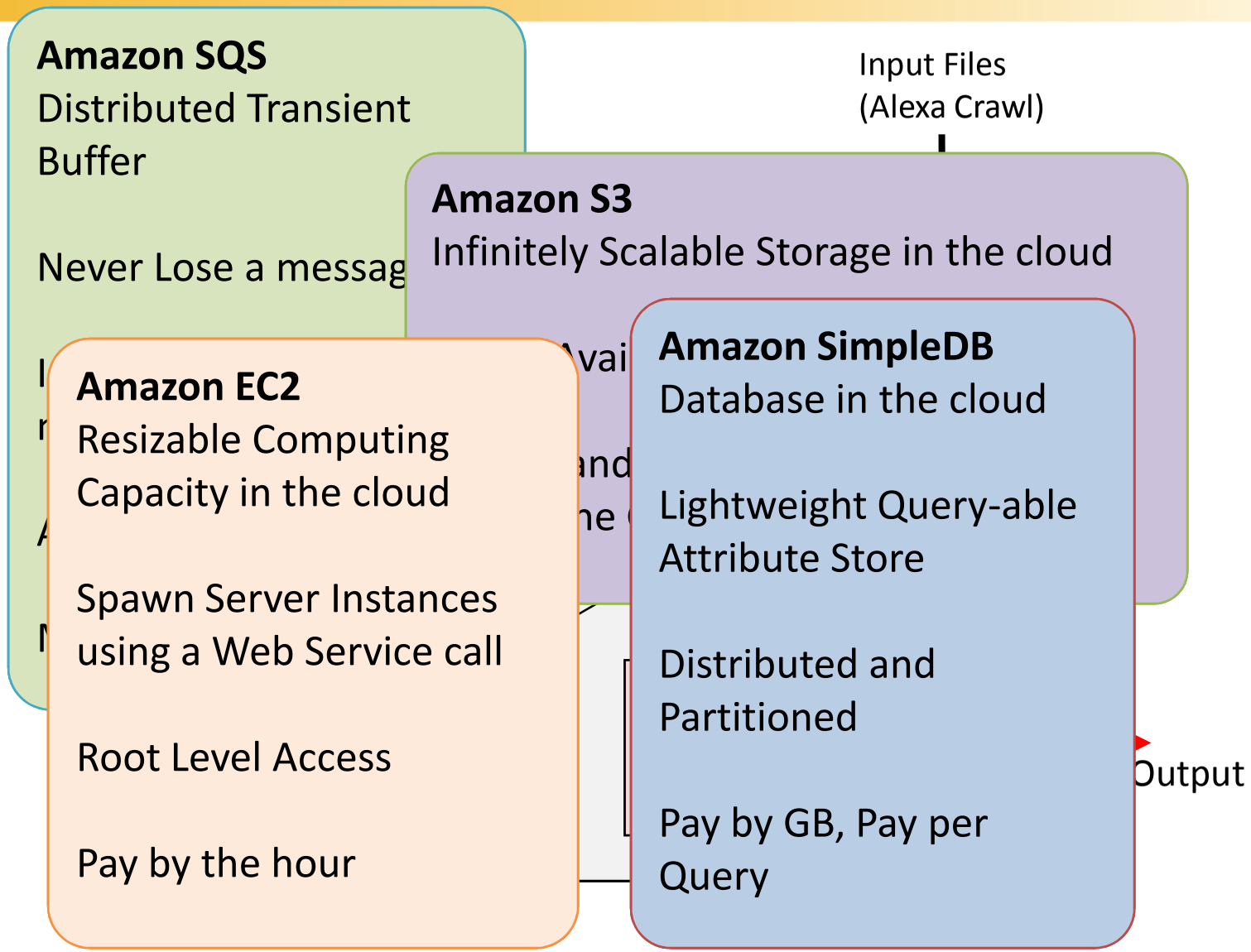
Examples of Patterns

- Source Code
  - int x = 40 + i
- Any thing with punctuation
  - “Hey!” he said, “Are you ok?”
- Case Sensitive
  - Function CallOrderController()
- Equations
  - f(x) = x^2
- Other Patterns
  - (dis)integration of life, Email Address

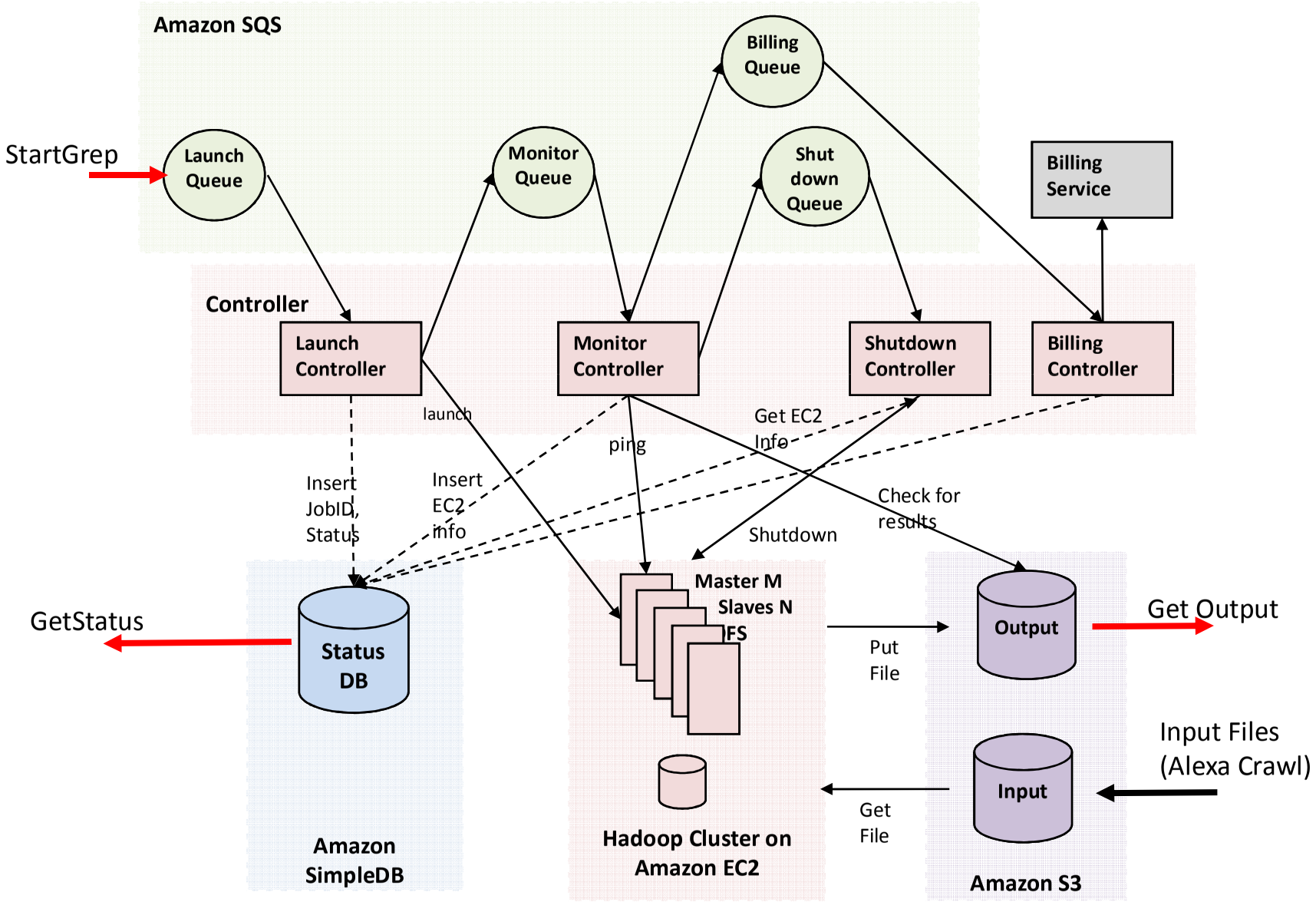
Zoom Level 1


Input dataset (List of Document Urls)
RegEx
Alexa GrepTheWeb Service
GetStatus
Subset of document URLs that matched the RegEx

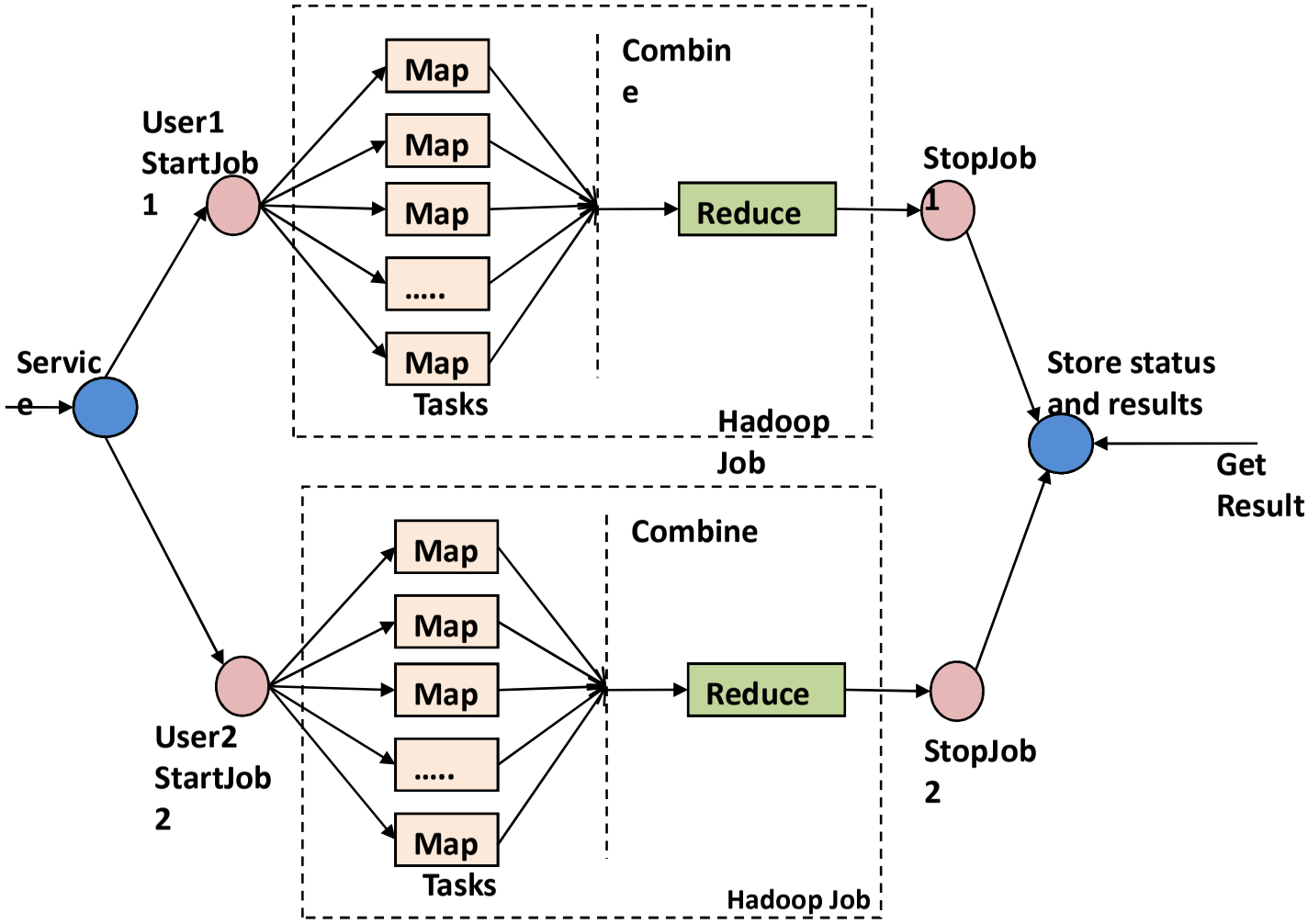
Zoom Level 2


Amazon SQS
Distributed Transient Buffer
Never Lose a messag
Input Files
(Alexa Crawl)
Amazon S3
Infinitely Scalable Storage in the cloud
Amazon SimpleDB
Database in the cloud
Lightweight Query-able Attribute Store
Distributed and Partitioned
Pay by GB, Pay per Query
Amazon EC2
Resizable Computing Capacity in the cloud
Spawn Server Instances using a Web Service call
Root Level Access
Pay by the hour
Output

Zoom Level 3


Amazon SQS
Billing Queue
StartGrep
Launch Queue
Monitor Queue
Shut down Queue
Billing Service
Controller
Launch Controller
Monitor Controller
Shutdown Controller
Billing Controller
launch
ping
Get EC2 Info
Insert JobID, Status
Insert EC2 info
Check for results
Shutdown
GetStatus
Status DB
Master M
Slaves N
FS
Put File
Output
Get Output
Input Files (Alexa Crawl)
Input
Get File
Amazon SimpleDB
Hadoop Cluster on Amazon EC2
Amazon S3

Zoom Level 4


Map
Map
Map
.....
Map
Tasks
Combin
e
Reduce
Hadoop
Job
User1
StartJob
1
StopJob
1
Servic
e
Store status
and results
Get
Result
Map
Map
Map
.....
Map
Tasks
Combine
Reduce
Hadoop Job
User2
StartJob
2
StopJob
2

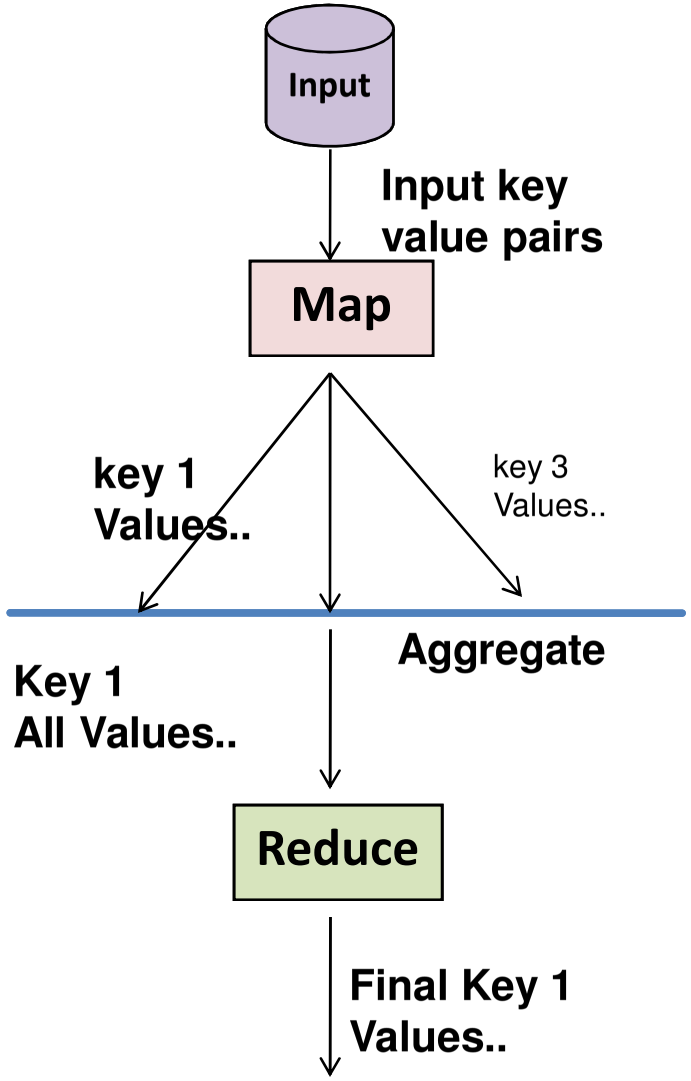
SideTrack: WordCount Example

MAPPER: For each input record, extract a set of key/value pairs that we care about the each record

```
"Hi Hadoop, Bye Hadoop"
            ↓
("Hi", 1), ("Hadoop", 1),
("Bye", 1), ("Hadoop", 1)
```

REDUCER: For each extracted key/value pair, combine it with other values that share the same key

```
("Hadoop", [1,1])
        ↓
  ("Hadoop", 2)
```



Source: Doug Cutting's Slide Deck on Hadoop

Zoom Level 5 (Hadoop MapReduce)

MAPPER: For each input record, extract a set of key/value pairs that we care about the each record

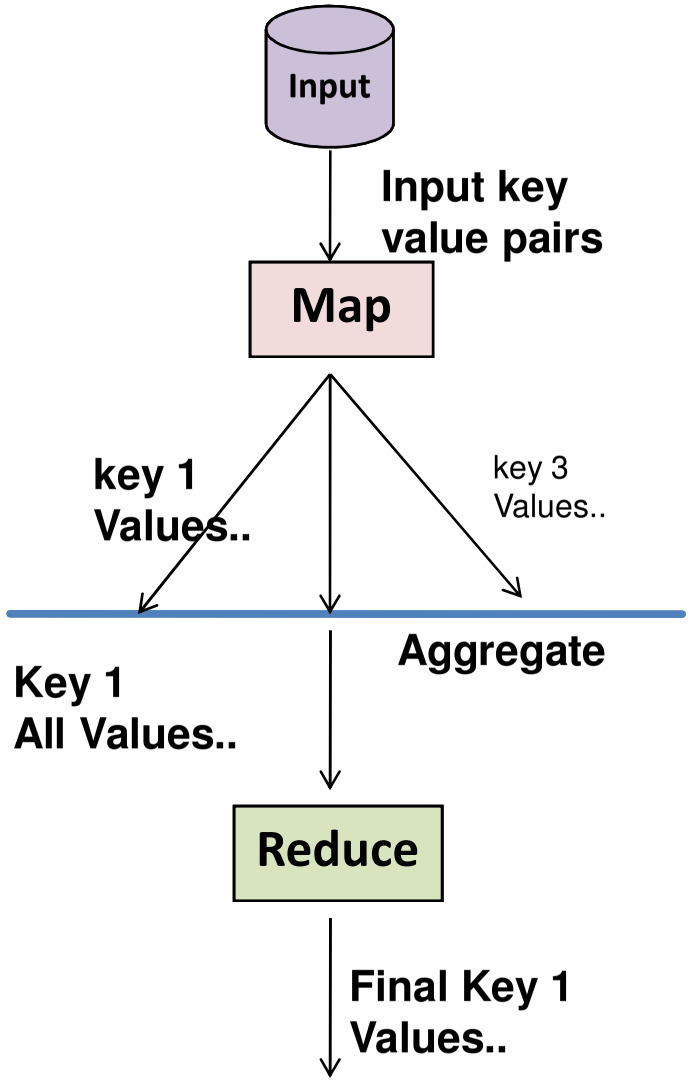
(LineNumber, s3pointer)

↓

(s3pointer, [matches])

REDUCER: For each extracted key/value pair, combine it with other values that share the same key

Identity Function



Source: Doug Cutting's Slide Deck on Hadoop

The Open Source Hadoop framework is giving developers the power to do some pretty extraordinary things.

The Open Source Hadoop framework ***on Amazon EC2/S3*** *is giving* ***every developer*** *the power to do some pretty extraordinary things.*

References

- Running Hadoop MapReduce on Amazon EC2 and Amazon S3

http://developer.amazonwebservices.com/connect/entry.jspa?externalID=873

- Hadoop on Amazon EC2 Step By Step Wiki

http://wiki.apache.org/hadoop/AmazonEC2

- Taking Massive Distributed Computing to the Common Man - Hadoop on Amazon EC2/S3

http://aws.typepad.com/aws/2008/02/taking-massive.html


Culture of Computing has changed
Economics of Computing has changed
Education in Computing has changed
Concepts in Computing has changed

Thank you!

Jinesh Varia
jvaria@amazon.com

http://aws.amazon.com /s3
/ec2
/sdb
/sqs
/forums
/resources
/blog