AWS for Data Analytics: Elastic MapReduce

By Jennifer Rubinovitz

@rubinovitz
j@rubinovitz.com

http://bit.ly/hadoopforhackers

Me

- Rutgers alumnus ‘13
- hackNY fellow ‘12 + hackNY mentor ‘13

Presently:

- C.S. M.S. student at Columbia working on Lean Workbench to quantify early stage startups

Outline

- What is MapReduce
- What is Hadoop
- What is Hadoop on AWS
- Example

Google’s Bluffs Data Center

Motivation: For This Talk

□ Can we democratize data science yet?

□ EMR is elastic and cheap

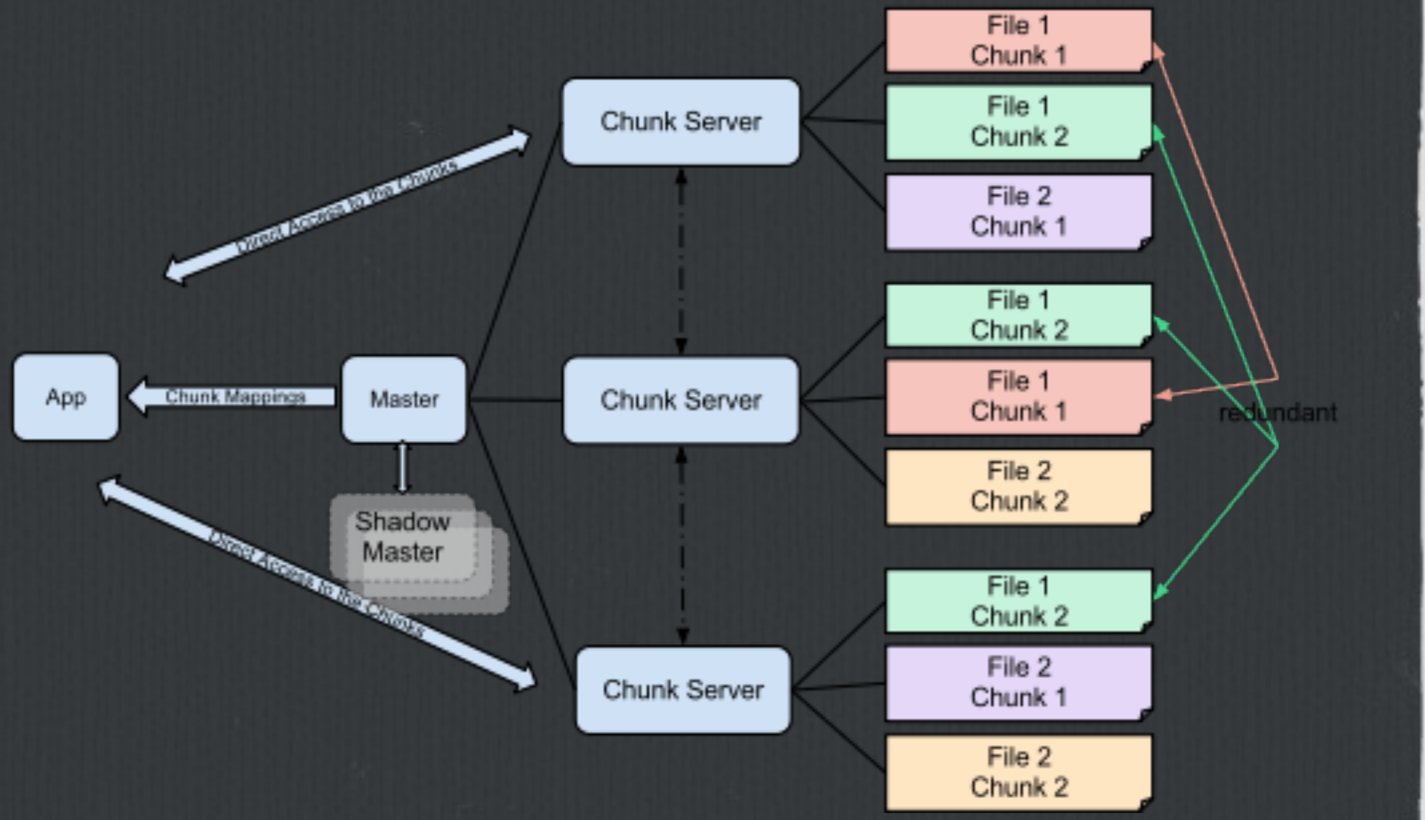
□ More online + offline resources than ever

Motivation: For MapReduce

□ You have data so big you need to parallelize processing (e.g. Google's Index of the World Wide Web)

□ Since you have so many nodes you need to assume there will always be failures

Solution: Google MapReduce and Google File System



Google File System

MapReduce

Two Separate Tasks:

Map - takes a set of data and breaks it down into tuples (key/value pairs) to be distributed to worker nodes.

Reduce - takes the output from the workers and aggregate the result.

MapReduce: Analogy

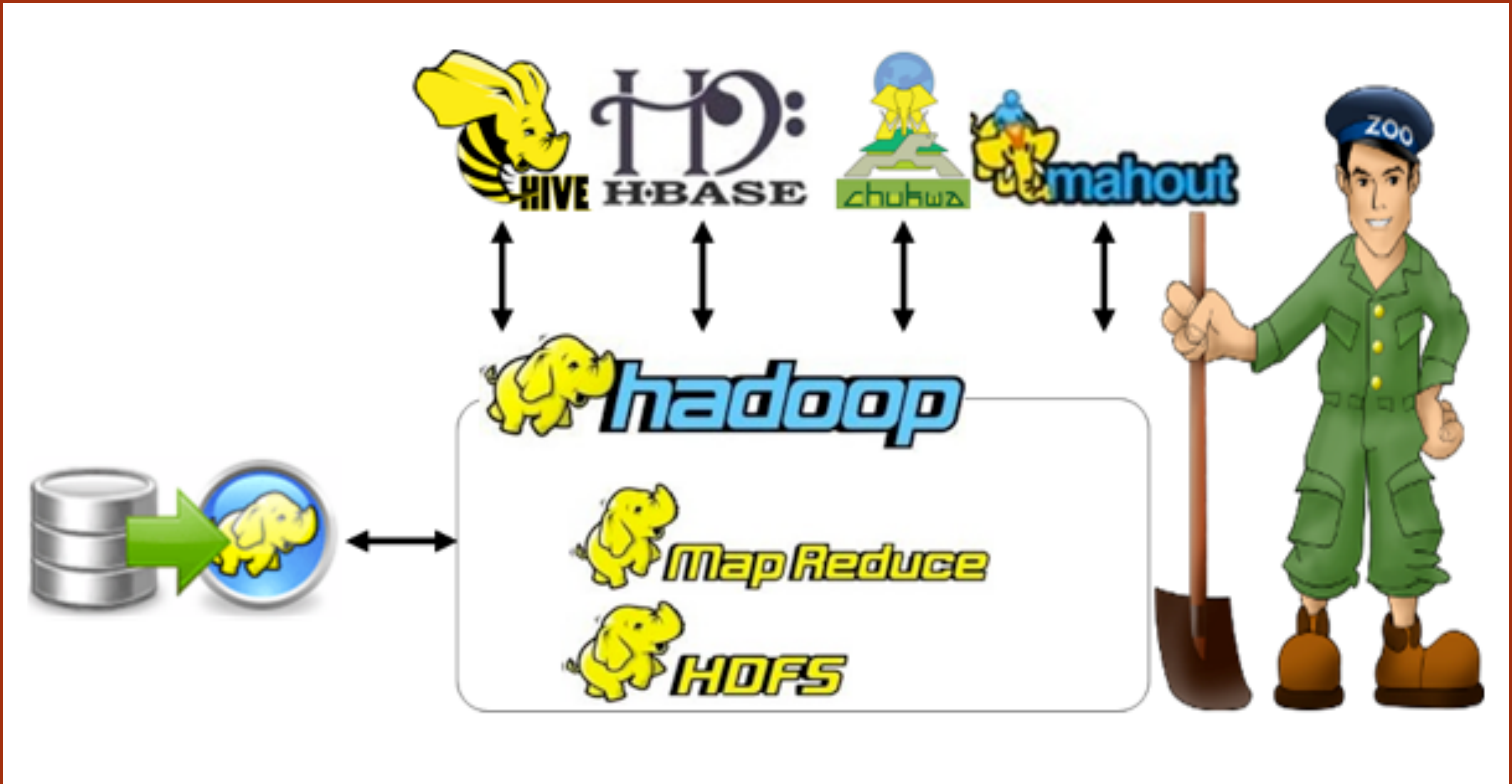
Roman Times Census:

□ Map: Census bureau would dispatch its people to each city in the empire. Each census taker in each city would be tasked to count the number of people in that city and then return their results to the capital city.

□ Reduce: Aggregate all results to a single count (sum of all cities) to determine the overall population of the empire.

Motivation: For Hadoop

- Google’s MapReduce and Filesystem are proprietary
- Hadoop is opensource software by Apache (you can use the software for free!)
- Not only is it opensource, the community is great!



Hadoop Ecosystem

Hadoop, Hive, HDFS, MapReduce, Mahout, HBase, Zookeeper, Chukwa, Pig

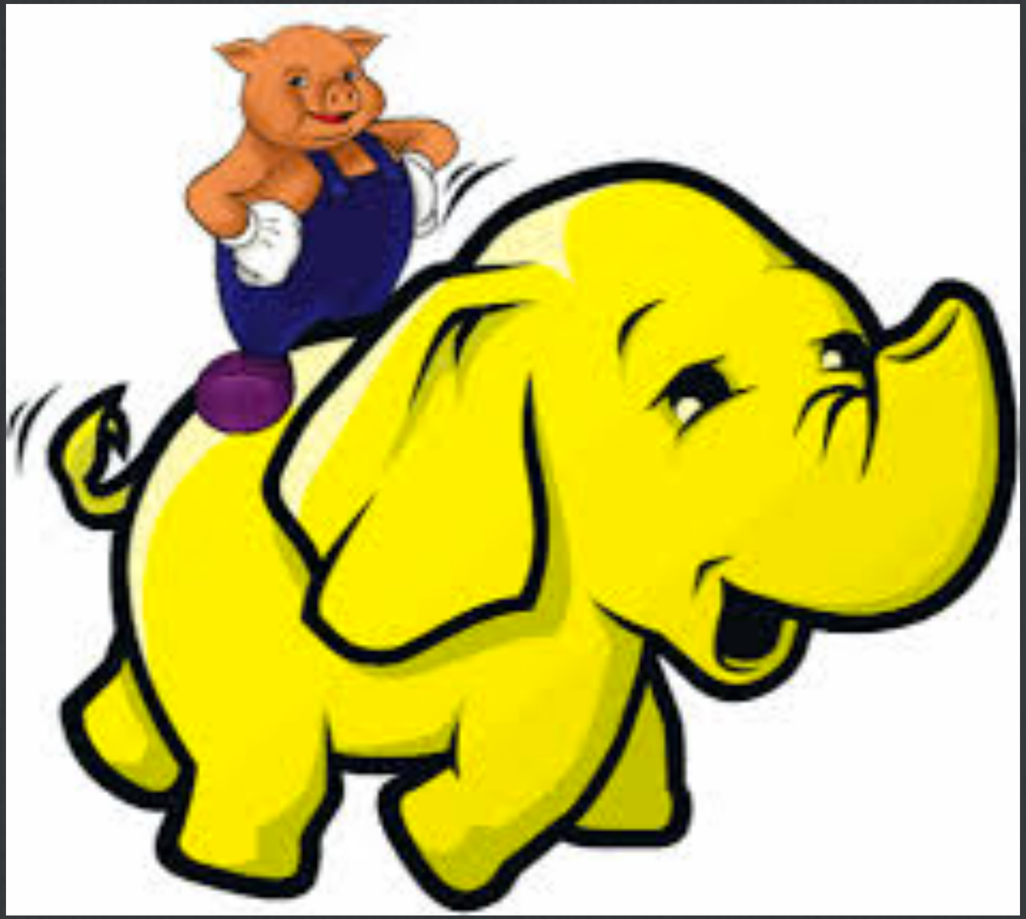
HBase

- distributed database modeled after Google's BigTable (which is built on top of the Google Filesystem)
- written in Java (but has wrappers for other languages)
- fault-tolerant
- good for sparse data

Hive

- SQL-like language for accessing HDFS

Pig



For data extraction like Hive, but has its own language: Pig Latin

Unlike Hive it can:

□ use lazy evaluation (delays the evaluation of an expression until its value is needed)

□ use ETL (Extract, Transform, Load)

□ store data at any point during a pipeline

□ declare execution plan

□ support pipeline splits

Other

- Chukwa - data collection system for monitoring large distributed systems
- Mahout - scalable machine learning libraries
- Zookeeper - service for maintaining clusters

Ways you may already use Hadoop

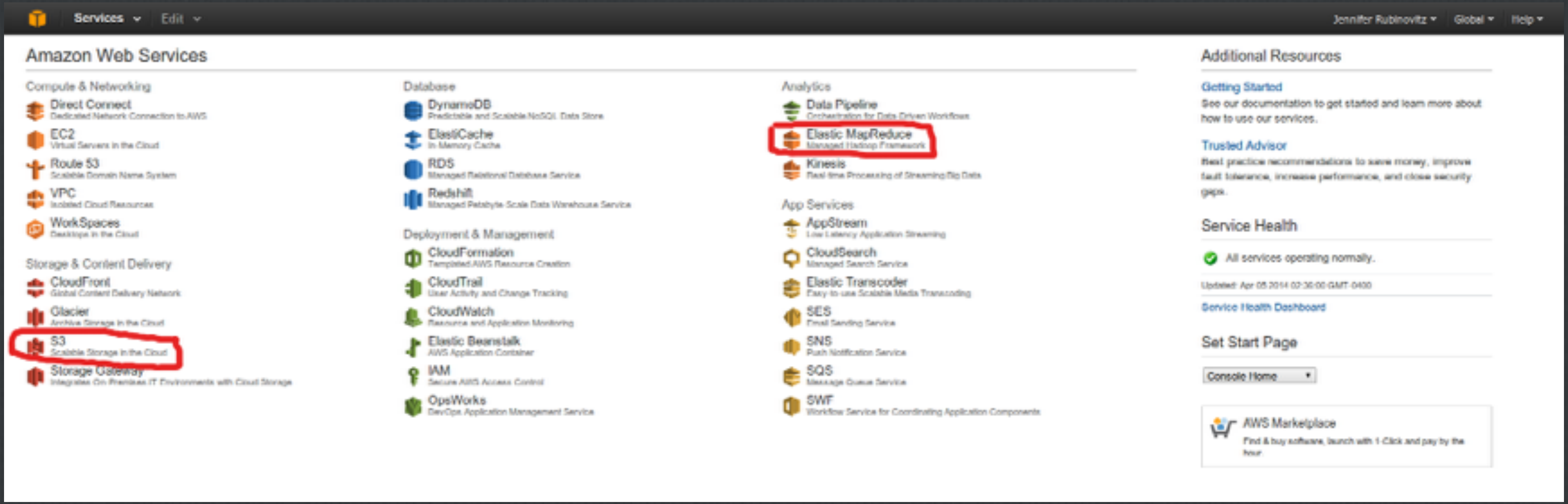
- □ HBase powers Facebook Messenger
- □ Yahoo! search is processed by Hadoop
- □ Ebay uses Hadoop for search optimization and research
- □ Hulu uses Hadoop and HBase for storage

And a ton of other products…

WHY: Hadoop on AWS

- Elastic
- Cheap
- Easier than doing DevOps yourself

AWS Basics



Amazon Web Services

Compute & Networking
- Direct Connect — Dedicated Network Connection to AWS
- EC2 — Virtual Servers in the Cloud
- Route 53 — Scalable Domain Name System
- VPC — Isolated Cloud Resources
- WorkSpaces — Desktops in the Cloud

Storage & Content Delivery
- CloudFront — Global Content Delivery Network
- Glacier — Archive Storage in the Cloud
- S3 — Scalable Storage in the Cloud
- Storage Gateway — Integrates On-Premises IT Environments with Cloud Storage

Database
- DynamoDB — Predictable and Scalable NoSQL Data Store
- ElastiCache — In-Memory Cache
- RDS — Managed Relational Database Service
- Redshift — Managed Petabyte-Scale Data Warehouse Service

Deployment & Management
- CloudFormation — Templated AWS Resource Creation
- CloudTrail — User Activity and Change Tracking
- CloudWatch — Resource and Application Monitoring
- Elastic Beanstalk — AWS Application Container
- IAM — Secure AWS Access Control
- OpsWorks — DevOps Application Management Service

Analytics
- Data Pipeline — Orchestration for Data-Driven Workflows
- Elastic MapReduce — Managed Hadoop Framework
- Kinesis — Real-time Processing of Streaming Big Data

App Services
- AppStream — Low Latency Application Streaming
- CloudSearch — Managed Search Service
- Elastic Transcoder — Easy-to-use Scalable Media Transcoding
- SES — Email Sending Service
- SNS — Push Notification Service
- SQS — Message Queue Service
- SWF — Workflow Service for Coordinating Application Components

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

Trusted Advisor
Best practice recommendations to save money, improve fault tolerance, increase performance, and close security gaps.

Service Health

✓ All services operating normally.

Updated: Apr 05 2014 02:36:00 GMT-0400

Service Health Dashboard

Set Start Page

Console Home ▾

AWS Marketplace
Find & buy software, launch with 1-Click and pay by the hour

□ Go to console.aws.amazon.com

HOW: Hadoop on AWS

- □ AWS Console GUI (www.console.aws.amazon.com)
- □ Ruby MapReduce Command Line Interface (http://aws.amazon.com/developertools/Elastic-MapReduce/2264): requires Ruby 1.8.7!!

Our Example

The Hello World of Hadoop is…

Word counting

Steps:

1. Write a mapper in Python

2. Put it into AWS S3

3. Launch an EMR instance

wordSplitter.py

```
#!/usr/bin/python
import sys
import re

def main(argv):
    pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
   for line in sys.stdin:
        for word in pattern.findall(line):
            print "LongValueSum:" + word.lower() + "\t" + "1"

if __name__ == "__main__":
    main(sys.argv)
```

Step 2: Setup a S3 Bucket For Storage

□ Go to https://console.aws.amazon.com/s3/

□ Click “Create Bucket”

□ Create a bucket to keep your data

Step 3: Setup Cluster

- Go to https://console.aws.amazon.com/elasticmapreduce/
- Click “Create Cluster”
- or in the CLI
- ./elastic-mapreduce —create —stream —mapper s3://<our-bucket>/wordSplitter.py —output s3://<our-bucket>/output —reducer aggregate

Step 4: Wait

□ Go running

□ Paint your nails

□ Read a book

□ Do other work? Naaa.

Example: Google Ngrams

- AWS has the 2 TB dataset of Ngrams in books over time
- We can use Hive to query them and find trends
- Dataset used for http://books.google.com/ngrams

Setup a hive instance

```
$ ./elastic-mapreduce —create —alive —hive-interactive

$ ./elastic-mapreduce —list <job-flow-id>

$ ./elastic-mapreduce —ssh <job-flow-id>
```

Setup Hive Tables

```
$ hive
$ set hive.base.inputformat=org.apache.hadoop.hive.ql.io.HiveInputFormat;
$ set mapred.min.split.size=134217728;
$ CREATE EXTERNAL TABLE english_1grams (
 gram string,
 year int,
 occurrences bigint,
 pages bigint,
 books bigint
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS SEQUENCEFILE
LOCATION 's3://datasets.elasticmapreduce/ngrams/books/20090715/eng-all/1gram/';
```

Normalize the Data

(convert to lowercase and ignore extraneous characters)

```
CREATE TABLE normalized (
 gram string,
 year int,
 occurrences bigint
);
INSERT OVERWRITE TABLE normalized
SELECT
 lower(gram),
 year,
 occurrences
FROM
 english_1grams
WHERE
 year >= 1890 AND
 gram REGEXP "^[A-Za-z+'-]+$";
```

Word-ratio by Decade

```
CREATE TABLE by_decade (

 gram string,

 decade int,

 ratio double

);
```

```
INSERT OVERWRITE TABLE by_decade

SELECT

 a.gram,

 b.decade,

 sum(a.occurrences) / b.total

FROM

 normalized a

JOIN (

 SELECT

  substr(year, 0, 3) as decade,

  sum(occurrences) as total

 FROM

  normalized

 GROUP BY

  substr(year, 0, 3)

) b

ON

 substr(a.year, 0, 3) = b.decade

GROUP BY

 a.gram,

 b.decade,

 b.total;
```

Results

1900

radium, ionization, automobiles, petrol, archivo, automobile, electrons, mukden, anopheles, marconi, botha, ladysmith, lhasa, boxers, suprema, aboord, rotor, turkes, wireless, conveyor, manchurian, erythrocytes, shoare, thirtie, kop, tuskegee, thorium, audiencia, bvo, arteriosclerosis

1910

cowperwood, britling, boches, montessori, venizelos, bolsheviki, salvarsan, photoplay, pacifists, joffre, petrograd, pacifist, bolshevism, airmen, kerensky, foch, boche, serbia, serbian, hindenburg, madero, serbians, bombing, ameen, anaphylaxis, aviators, syndicalism, aviator, biplane, taxi

…

1930

dollfuss, goebbels, manchukuo, hitler, sudeten, hitler's, rearmament, nazis, wpa, nazi, nra, manchoukuo totalitarian, pwa, tva, stalin's, peiping, homeroom, kulaks, stalin, devaluation, bta, carotene, broadcasts, corporative, comintern, ergosterol, reichswehr, ussr, businessmen

…

2000

bibliobazaar, itunes, cengage, qaeda, wsdl, aspx, xslt, actionscript, xpath, sharepoint, blogs, easyread, ipod, xhtml, blog, rfid, google, writeline, proteomics, bluetooth, voip, microarray, mysql, microarrays, putin, dreamweaver, dvds, ejb, xml, osama

Next Steps

- □ Go to www.github.com/rubinovitz/hackny-masters-hadoop for more examples and reference
- □ Questions?