Big Data

Facebook Friends Data on Amazon Elastic Cloud
Agenda

Cloud Computing
• Taxonomy
• Google Cloud
• Amazon Cloud
• Comparing Amazon and Google - BATTLE IS ON
• Amazon EC2 detailed study

Big Data Processing
• Our data
• Transforming our data
• Big Data Processing System - APACHE HIVE
• Data mining - FINALLY !!
Cloud Computing
**Cloud Computing Service Models**

- **Software as a Service (SaaS)**
  
  How applications are delivered (i.e. Salesforce.com, Gmail). Software resides on the software providers servers or a 3rd party host.

- **Platform as a Service (PaaS)**

  Where applications are deployed. Base runtime environments (i.e. database, application server).

- **Infrastructure as a Service (IaaS)**

  Where runtime environments are deployed. Base layer for hardware and networking (i.e. Servers, Storage, Virtual Machine).
Google Apps, Salesforce.com, Netsuite, Yahoo Mail, ...

Google App Engine, Windows Azure, AWS Elastic Beanstalk, ...

Amazon EC2, RackSpace, Google Compute Engine, ...
Google Cloud Platform is a set of services that enables developers to build, test and deploy applications on Google’s reliable infrastructure.
Google App Engine - April 2008

- Platform as a Service.
- For developing and hosting web applications in Google managed data centers on multiple servers.
- Scalable cloud solutions for webapps and mobile backend.
- Supports for different runtime environment based on your needs.
Google App Engine (cont.)

The Managed VM hosting environment lets you run App Engine Applications on configurable Compute Engine Virtual Machines (VMs)
Google Compute Engine - 2012

- Infrastructure as a Service.

- Service that provides virtual machines that run on Google infrastructure.

- Capabilities of GCE:
  - Create virtual machines with variety of configurations.
  - Maintain and store data in block storages.
  - Manage network access to your virtual machines.
  - Use variety of tools and OAuth authentication to manage your virtual machines.
Google Compute Engine (cont.)

- Each virtual machine is considered an instance resource.
- Each instruction resource uses other resources such as Disk resources, Network resources, Image resources and so on.
- Resources divided into global, regional and zonal resource plane.
- This provides higher degree of failure independence.
Amazon Web Services - AWS

- AWS is a evolving cloud computing platform provided by Amazon.com.
- It was first introduced in 2006.
- To ensure robustness of the system AWS servers are installed at 7 different locations - Eastern USA, Western USA, Brazil, Ireland, Singapore, Japan, and Australia.
- AWS, in total, offers more 36 services.
Amazon Web Services - AWS

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

### Storage & CDN
- S3: Scalable Storage in the Cloud
- EBS: Networked Attached Block Device
- CloudFront: Global Content Delivery Network
- Glacier: Archive Storage in the Cloud
- Storage Gateway: Integrates On-Premises IT with Cloud Storage
- Import Export: Ship Large Datasets

### Cross-Service
- Support: Phone & email fast-response 24X7 Support
- Marketplace: Buy and sell Software and Apps
- Management Console: UI to manage AWS services
- SDKs, IDE kits and CLIs: Develop, integrate and manage services

### Analytics
- Elastic MapReduce: Managed Hadoop Framework
- Kinesis: Real-Time Data Stream Processing
- Data Pipeline: Orchestration for Data-Driven Workflows

### Compute & Networking
- EC2: Virtual Servers in the Cloud
- VPC: Virtual Secure Network
- ELB: Load Balancing Service
- WorkSpaces: Virtual Desktops in the cloud
- Auto Scaling: Automatically scale up and down
- DirectConnect: Dedicated Network Connection to AWS
- Route 53: Scalable Domain Name System

### Deployment & Management
- CloudFormation: Templated AWS Resource Creation
- CloudWatch: Resource and Application Monitoring
- Elastic Beanstalk: AWS Application Container
- IAM: Secure AWS Access Control
- CloudTrail: User Activity Logging
- OpsWorks: DevOps Application Management Service
- CloudHSM: Hardware-based key storage for compliance

### App Services
- 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 App Components
- AppStream: Low-latency Application Streaming

### AWS Global Physical Infrastructure
(Geographical Regions, Availability Zones, Edge Locations)
AWS - Elastic Compute Cloud (EC2)

- EC2 is a web service that provides resizable compute capacity in the cloud.
- Allows you to rent virtual computers to run your personal application.
- Using EC2 you can launch various operating systems, load custom application environments, and manage network permissions.
EC2 - Benefits

• Elastic Web-Scale Computing
• Flexible Cloud Hosting Services
• Usable with other Amazon Web Services
• Reliable
• Inexpensive
  • On-Demand Instances
  • Reserved Instances
  • Spot Instances
Google Compute Engine VS AWS EC2

• Load Balancing
  • AWS’ ELB cannot deal with an unexpected rise of traffic.
  • To enable ELB to handle unexpected rise of traffic end user has to do some configuration work.
  • Google has “Native Load-Balancing Technology” which helps you spread the traffic on pool of instances.

• Stability
  • Since GCE was officially made available in 2013 it still lacks stability as compared to EC2 which was introduced in 2006.
Google Compute Engine VS AWS EC2

- **Flexibility**
  - **AWS** offers lots of stable services for storage, computing, databases, networking and many more.
  - On the other hand GCE, being new player in the market still has to work on developing various different services.

- **Performance and Costing**
  - Due to GCE’s pre existing infrastructure and efficient implementation, performance and costing is better than AWS.
How to Setup Amazon EC2 Instance?
Select AMI

Step 1: Choose an Amazon Machine Image (AMI)

- Red Hat Enterprise Linux 6.5 (HVM) - ami-11125e21
  - Red Hat Enterprise Linux version 6.5 (HVM), EBS-backed
  - Root device type: ebs
  - Virtualization type: hvm
  - 64-bit

- SuSE Linux Enterprise Server 11 SP3 (HVM), SSD Volume Type - ami-3b0420b
  - SuSE Linux Enterprise Server 11 Service Pack 3 (HVM), EBS General Purpose (SSD) Volume Type. Nvidia driver installs automatically during startup for GPU instances.
  - Root device type: ebs
  - Virtualization type: hvm
  - 64-bit

- Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-3d50120d
  - Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services).
  - Root device type: ebs
  - Virtualization type: hvm
  - 64-bit

- Microsoft Windows Server 2012 R2 Base - ami-d38d0ec3
  - Microsoft Windows 2012 R2 Standard edition with 64-bit architecture. [English]
  - Root device type: ebs
  - Virtualization type: hvm
  - 64-bit

Are you launching a database instance? Try Amazon RDS.
Select Type of Instance

Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. Learn more about instance types and how they can meet your computing needs.

<table>
<thead>
<tr>
<th>Family</th>
<th>Type</th>
<th>vCPUs</th>
<th>Memory (GiB)</th>
<th>Instance Storage (GiB)</th>
<th>EBS-Optimized Available</th>
<th>Network Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>General purpose</td>
<td>t2.micro</td>
<td>1</td>
<td>1</td>
<td>EBS only</td>
<td>-</td>
<td>Low to Moderate</td>
</tr>
<tr>
<td>General purpose</td>
<td>t2.small</td>
<td>1</td>
<td>2</td>
<td>EBS only</td>
<td>-</td>
<td>Low to Moderate</td>
</tr>
<tr>
<td>General purpose</td>
<td>t2.medium</td>
<td>2</td>
<td>4</td>
<td>EBS only</td>
<td>-</td>
<td>Low to Moderate</td>
</tr>
<tr>
<td>General purpose</td>
<td>m3.medium</td>
<td>1</td>
<td>3.75</td>
<td>1 x 4 (SSD)</td>
<td>-</td>
<td>Moderate</td>
</tr>
<tr>
<td>General purpose</td>
<td>m3.large</td>
<td>2</td>
<td>7.5</td>
<td>1 x 32 (SSD)</td>
<td>-</td>
<td>Moderate</td>
</tr>
</tbody>
</table>
Configure the Instance

Step 3: Configure Instance Details
Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot Instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

- **Number of instances**: 1
- **Purchasing option**: None
- **Network**: vpc-140t75f1 (172.31.0.0/16) (default)
- **Subnet**: subnet-24029830 (172.31.16.0/29) Default in us-west-2a
  
  4091 IP Addresses available
- **Auto-assign Public IP**: Use subnet setting (Enable)
- **IAM role**: None
- **Shutdown behavior**: Stop
- **Enable termination protection**: None
- **Monitoring**: None

Additional CloudWatch detailed monitoring

Additional charges apply.
Add Storage

Step 4: Add Storage
Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. Learn more about storage options in Amazon EC2.

<table>
<thead>
<tr>
<th>Type</th>
<th>Device</th>
<th>Snapshot</th>
<th>Size (GB)</th>
<th>Volume Type</th>
<th>IOPS</th>
<th>Delete on Termination</th>
<th>Encrypted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root</td>
<td>/dev/sda1</td>
<td>snap-ddd48814</td>
<td>15</td>
<td>General Purpose (SSD)</td>
<td>45 / 3000</td>
<td>Not Encrypted</td>
<td></td>
</tr>
</tbody>
</table>

Add New Volume

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. Learn more about free usage tier eligibility and usage restrictions.
Assign Tag to Instance

Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. Learn more about storage options in Amazon EC2.

<table>
<thead>
<tr>
<th>Type</th>
<th>Device</th>
<th>Snapshot</th>
<th>Size (GiB)</th>
<th>Volume Type</th>
<th>IOPS</th>
<th>Delete on Termination</th>
<th>Encrypted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root</td>
<td>/dev/sda1</td>
<td>snap-ddd48814</td>
<td>15</td>
<td>General Purpose (SSD)</td>
<td>45 / 3000</td>
<td>0</td>
<td>Not Encrypted</td>
</tr>
</tbody>
</table>

Add New Volume

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. Learn more about free usage tier eligibility and usage restrictions.
Configure the Security Groups and Launch the Instance

Step 7: Review Instance Launch

- **Instance Type**
  - t2.micro
    - ECUs: Variable
    - vCPUs: 1
    - Memory (GiB): 1
    - Instance Storage (GB): EBS only
    - EBS-Optimized Available: -
    - Network Performance: Low to Moderate

- **Security Groups**
  - Security group name: launch-wizard-1
    - Description: Created 2014-10-22T01:18:21.689-04:00
  - Type: SSH
    - Protocol: TCP
    - Port Range: 22
    - Source: 0.0.0.0/0

- **Instance Details**

- **Storage**
Create .pem File

Select an existing key pair or create a new key pair

A key pair consists of a public key that AWS stores, and a private key file that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about managing access key pairs from a public AMI.

- Create a new key pair
- Proceed without a key pair

Key pair name

Download Key Pair

You have to download the private key file (*.pem file) before you can continue. Store it in a secure and accessible location. You will not be able to download the file again after it's created.

Cancel  Launch Instances
Big Data Processing
Our Data (cont.)

- Data Source: Facebook - friends data

- Downloading data:
  - Initial approach: Graph Explorer API from Facebook
    The new version of GRAPH API from Facebook includes restriction of friend data to friends that have connected to the client application.
    - Limitations for friends data:
      - Restriction on fetching complete friend list of user.
      - Only available friends data is through taggable and invitable friends.
      - Taggable and Invitable friends only give user name and tag id.
  - Alternative approach:
    - Give me my Data (Popular Facebook Application):
      It helps users to export their data out of Facebook for reuse in visualizations, archives or any possible method of digital story telling.
Our data (cont.)

- Result obtained in semi structured XML Format.
- One chunk of data for single friend is shown in the next slide.
<friend>
    <uid>513084950</uid>
    <first_name>Sachin</first_name>
    <last_name>Lathia</last_name>
    <pic_small>https://fbcdn-profile-a.akamaihd.net/hprofile-ak-xap1/v/t1.0-302/1/p50x50/30637_444227474950_5590897_n.jpg?oh=b6edd0f259d11c531d6fddd0fb51f6268&amp;oe=54F088B6&amp;__gda__=1423436300_325c2ac2ffeacb14d615b7da0e19150</pic_small>
    <pic_big>https://fbcdn-profile-a.akamaihd.net/hprofile-ak-xap1/v/t1.0-302/1/p200x200/30637_444227474950_5590897_n.jpg?oh=7ca59ed897a215d3773a1e66b3fe0cb&amp;oe=54D5555E&amp;__gda__=1420532842_113c096456741d9f5f49987f04f7</pic_big>
    <pic_square>https://fbcdn-profile-a.akamaihd.net/hprofile-ak-xap1/v/t1.0-302/1/c47.29.333.333/s50x50/30637_444227474950_5590897_n.jpg?oh=83669bc2dce85c58803e482561ba44&amp;oe=54F831EC&amp;__gda__=1424284733_df63115d18d377377809e078f0130b</pic_square>
    <pic>https://fbcdn-profile-a.akamaihd.net/hprofile-ak-xap1/v/t1.0-302/1/p100x100/30637_444227474950_5590897_n.jpg?oh=e9c27a15e677e26e315bb2819886aead&amp;oe=54E75828&amp;__gda__=1425226524_9f052b9a75867e4ce5e1895c71fba</pic>
    <profile_update_time>2013-10-28T23:45:08-07:00</profile_update_time>
    <timezone></timezone>
    <religion></religion>
    <birthday></birthday>
    <birthday_date></birthday_date>
    <sex>male</sex>
    <hometown_location>
        <city>Rajkot</city>
        <state>Gujarat</state>
        <country>India</country>
        <zip></zip>
        <latitude>22.3</latitude>
        <longitude>70.7833</longitude>
        <id>110832712274056</id>
        <name>Rajkot, Gujarat</name>
    </hometown_location>
    <current_location>
        <city>Vadodara</city>
        <state>Gujarat</state>
        <country>India</country>
        <zip></zip>
        <latitude>22.3</latitude>
        <longitude>73.2</longitude>
        <id>106280092741214</id>
        <name>Vadodara, Gujarat, India</name>
    </current_location>
    <affiliations></affiliations>
    <meeting_for></meeting_for>
    <meeting_sex></meeting_sex>
    <relationship_status>Married</relationship_status>
    <significant_other_id>10000184048305</significant_other_id>
    <political></political>
    <activities>Counselling</activities>
    <is_app_user></is_app_user>
    <locale>en_GB</locale>
    <profile_url>https://www.facebook.com/sachin.lathia</profile_url>
    <website></website>
    <contact_email></contact_email>
    <email></email>
</friend>
Transforming Data

- The XML data was then processed using PHP script to a legible array, an intermediate format that is required for a final transformation.

```php
Array
(
    [0] => Array
        (        
            [uid] => 513084950
            [first_name] => Sachin
            [last_name] => Lathia
            [name] => Sachin Lathia
            [profile_update_time] => 2013-10-29 02:45:08
            [birthday] =>
            [sex] => male
            [relationship_status] => Married
            [activities] => Counselling
            [locale] => en_GB
            [profile_url] => https://www.facebook.com/sachin.lathia
            [home_location_city] => Rajkot
            [home_location_state] => Gujarat
            [home_location_country] => India
            [current_location_city] => Vadodara
            [current_location_state] => Gujarat
            [current_location_country] => India
        )
)```
Finally, the intermediate array is converted to CSV, a widely used and simple way to upload data on Hive.
Why HIVE?

- Capable of fulfilling data warehousing needs.
- Easy to use SQL like query language.
- Query execution via Map reduce.
- A mechanism to impose structure on variety of data formats (casting).
Apache Hive is a data warehouse software that facilitates querying and managing large datasets residing in distributed storage.

- It uses Hadoop (HDFS file structure) for storage and MapReduce for execution.
HIVE - Benefits

• Uses HIVEQL for querying similar to SQL
• Extensible - Add UDTs and UDF
• Performance
HIVE - Data Types

• Primitive types :
  • Integers : TINYINT, SMALLINT, INT, BIGINT
  • Boolean : BOOLEAN
  • Floating point numbers : FLOAT, DOUBLE
  • String : STRING

• Complex types :
  • Structs : {a INT; b INT}
  • Maps : M['group']
  • Arrays : ['a', 'b', 'c'], A[1] returns 'b'
HIVE - Data Models

- Tables - Each table has corresponding directory in HDFS
- Partitions - Nested sub-directories in HDFS for each partition column values
- Buckets - Split data based on hash of a column - mainly because of parallelism
CREATE TABLE sample (foo INT, bar STRING) PARTITIONED BY (ds STRING);

SHOW TABLES '.*s';

DESCRIBE sample;

ALTER TABLE sample ADD COLUMNS (new_col INT);

DROP TABLE sample;
HIVE - Create Table

• CREATE TABLE user_friends (uid INT, first_name VARCHAR(1024), last_name VARCHAR(1024), name STRING, profile_update_time STRING, birthday STRING, sex VARCHAR(1024), relationship_status VARCHAR(1024), activities VARCHAR(1024), locale VARCHAR(1024), profile_url VARCHAR(1024), home_location_city VARCHAR(1024), home_location_state VARCHAR(1024), home_location_country VARCHAR(1024), current_location_city VARCHAR(1024), current_location_state VARCHAR(1024), current_location_country VARCHAR(1024)) ROW FORMAT DELIMITED FIELDS TERMINATED BY ', ' STORED AS TEXTFILE;
HIVE - Load Data

- LOAD DATA INPATH ‘/user/Ubuntu/books/file.csv’ OVERWRITE INTO TABLE user_friends;
HIVE - Select Data

- Select * from user_friends;

```sql
hive> select * from user_friends;

Time taken: 0.056 seconds, Fetched: 1 row(s)
```
Data mining

Display information of all friends:

Query: Select * from user_friends;

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Birthday</th>
<th>Gender</th>
<th>Education</th>
<th>Hobbies</th>
<th>Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>513004956</td>
<td>Sachin Kathi</td>
<td>1997-04-28</td>
<td>Male</td>
<td>B.Tech Electronics</td>
<td>Reading, Music</td>
<td>Hiking, Yoga</td>
</tr>
<tr>
<td>515666322</td>
<td>Ricky Sutaria</td>
<td>1991-01-05</td>
<td>Male</td>
<td>B.Tech Electronics</td>
<td>Reading, Music</td>
<td>Hiking, Yoga</td>
</tr>
</tbody>
</table>
Data mining (cont.)

Find the number of friends who live in India:

Query:

```
SELECT COUNT(first_name) FROM user_friends WHERE home_location_country = 'India';
```

```
hive> SELECT COUNT(first_name) FROM user_friends WHERE home_location_country = 'India';


Starting Job = job_201411181846_0001

MapReduce jobs launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.36 sec HDFS Read: 239066 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 360 msec

Time taken: 24.25 seconds, fetched: 1 row(s)
```
Data mining (cont.)

Find the number of friends born after 1991:

Query:
SELECT COUNT(first_name) FROM user_friends WHERE to_date(birthday) >= "1991-01-01";
Find the number of friends with relationship status single:

**Query:**

```sql
SELECT COUNT(first_name) FROM user_friends WHERE relationship_status = 'Single';
```

```bash
Time taken: 0.066 seconds, Fetched: 5 row(s)
hive> SELECT COUNT(first_name) FROM user_friends WHERE relationship_status = 'Single';
Query ID = ubuntu_20141118191818_98ead496-cb3b-4b74-bbf8-150422f885e8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Kill Command = /home/ubuntu/hadoop/sbin/hadoop job -kill job_20141118191814_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2014-11-18 19:18:38,533 Stage-1 map = 0%, reduce = 0%
2014-11-18 19:18:42,523 Stage-1 map = 100%, reduce = 0%
Cumulative CPU 1.3 sec
2014-11-18 19:18:51,947 Stage-1 map = 100%, reduce = 100%
Cumulative CPU 2.37 sec
MapReduce Total cumulative CPU time: 2 seconds 370 msec
Ended Job = job_20141118191814_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.37 sec HDFS Read: 239066 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 370 msec
OK
393
Time taken: 20.414 seconds, Fetched: 1 row(s)
hive>
```