Large-scale Data Processing on the Cloud

MTAT.08.036
Lecture 1: Data analytics in the cloud

Satish Srirama
satish.srirama@ut.ee
Course Purpose

• Introduce cloud computing concepts
• Introduce data analytics in the cloud
• Introduce
  – Distributed computing algorithms such as MapReduce
  – BigData solutions such as Pig and Spark
  – NoSQL
• Glance of research at Mobile & Cloud Lab in cloud computing domain

• http://courses.cs.ut.ee/2017/LDPC/
Course Schedule

• Lectures: (Weeks 1-8)
  – Wed. 10.15 - 12.00 (Ülikooli 17 - 218)

• Practice sessions: (Weeks 2-9)
  – Tue. 10.15 - 12.00 (J. Liivi 2 - 003)
Related Courses

• **MTAT.08.027** Basics of Cloud Computing (3 ECTS)
  – Spring 2018

• **MTAT.03.280** Mobile and Cloud Computing Seminar (3 ECTS)
  – Thu. 14.15 - 16.00, Ülikooli 17 - 220

• **MTAT.03.266** Mobile Application Development Projects (3 ECTS)
  – Tue. 10.15 - 12.00, Ülikooli 17 - 219
Questions

• Is everyone comfortable with data structures?
• How comfortable you are with algorithms?
• How comfortable you are with programming?
  – Java?
    • External APIs?
  – Web programming
Grading

• Written exam – 50%
• Labs – 45%
  – 8 lab exercises
  – You must submit lab exercise solutions by Monday night of the week after (for full score)
    • Late submission for a week (80%)
    • Late submission for 2 weeks (50%) – Hard deadlines will be mentioned later
• Active participation in the lectures (Max 5%)

• To pass the course
  – You need to score at least 50% in each of the subsections
  – You need to score at least 50% in the total
Course schedule

• 06.09 Data analytics in the cloud
• 13.09 MapReduce
• 20.09 MapReduce algorithms
• 27.09 Information Retrieval with MapReduce
• 04.10 Apache Pig
• 11.10 Apache Spark
• 18.10 DataFrame and SQL abstractions
• 25.10 NoSQL Databases
• 01.11 Examination 1
Course schedule - continued

• Labs
  
  12.09 Utilizing Cloud resources  
  19.09 MapReduce - Basics  
  26.09 Data analysis with MapReduce  
  03.10 Information Retrieval with MapReduce  
  10.10 Processing data with Apache Pig  
  17.10 Processing data with Apache Spark  
  24.10 DataFrames in Spark  
  31.10 Utilizing NoSQL databases
Reference Books

• Mastering Cloud Computing: Foundations and Applications Programming
  Authors: Rajkumar Buyya, Christian Vecchiola, S.Thamarai Selvi

• Data-Intensive Text Processing with MapReduce
  Authors: Jimmy Lin and Chris Dyer

Reference Papers


Lecture 1

DATA ANALYTICS IN THE CLOUD
“It’s nothing new...”

“...we’ve redefined Cloud Computing to include everything that we already do... I don’t understand what we would do differently... other than change the wording of some of our ads.”

Larry Ellison, CEO, Oracle (Wall Street Journal, Sept. 26, 2008)

“It’s a trap”

“It’s worse than stupidity: it’s marketing hype. Somebody is saying this is inevitable—and whenever you hear that, it’s very likely to be a set of businesses campaigning to make it true.”


“No. It’s not true! Everyone thinks it is something else...”
What is Cloud Computing?

• Computing as a utility
  – Utility services e.g. water, electricity, gas etc.
  – Consumers pay based on their usage

1969 – Leonard Kleinrock, ARPANET project

“As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of ‘computer utilities’, which, like present electric and telephone utilities, will service individual homes and offices across the country”
Timeline

Grid Computing
- Solving large problems with parallel computing
- Made mainstream by Globus Alliance

Utility Computing
- Offering computing resources as a metered service
- Introduced in late 1990s

Software as a Service
- Network-based subscriptions to applications
- Gained momentum in 2001

Cloud Computing
- Next-Generation Internet computing
- Next-Generation Data Centers

https://www.slideshare.net/BijitGhosh1/presentation-on-cloud-computing-15257201
Clouds - Why Now (not then)?

• Experience with very large datacenters
  – Unprecedented economies of scale
  – Transfer of risk

• Technology factors
  – Pervasive broadband Internet
  – Maturity in Virtualization Technology

• Business factors
  – Minimal capital expenditure
  – Pay-as-you-go billing model
Virtualization

• Virtualization techniques are the basis of the cloud computing
• Virtualization technologies partition hardware and thus provide flexible and scalable computing platforms
• Virtual machine techniques
  – VMware and Xen
  – OpenNebula
  – Amazon EC2
Containers

• Use kernel of the host operating system instead of a hypervisor
  – Linux namespace
• Lightweight
  – No hypervisor overhead
  – Each Container acquires only required resources
• Fast start-up/ Bootup
  – Starting a container is faster than booting new OS or spinning up a new VM
• Performance
  – Near native performance
• Different container frameworks
  – LXC, Docker, Linux VServer, OpenVZ
  – Docker is an Open platform [https://www.docker.com/]
Cloud Computing - Characteristics

• Illusion of infinite resources
• No up-front cost
• Fine-grained billing (e.g. hourly)

• Gartner: “Cloud computing is a style of computing where massively scalable IT-related capabilities are provided ‘as a service’ across the Internet to multiple external customers”
Cloud Computing - Services

• **Software as a Service – SaaS**
  – A way to access applications hosted on the web through your web browser

• **Platform as a Service – PaaS**
  – Provides a computing platform and a solution stack (e.g. LAMP) as a service

• **Infrastructure as a Service – IaaS**
  – Use of commodity computers, distributed across Internet, to perform parallel processing, distributed storage, indexing and mining of data
  – Virtualization

  - **SaaS**
    - Facebook, Flikr, Myspace.com, Google maps API, Gmail
  - **PaaS**
    - Google App Engine, Force.com, Hadoop, Azure, Heroku, etc.
  - **IaaS**
    - Amazon EC2, Rackspace, GoGrid, SciCloud, etc.
Cloud Computing - Themes

• Massively scalable
• On-demand & dynamic
• Only use what you need - Elastic
  – No upfront commitments, use on short term basis
• Accessible via Internet, location independent
• Transparent
  – Complexity concealed from users, virtualized, abstracted
• Service oriented
  – Easy to use SLAs

SLA – Service Level Agreement
Cloud Models

• Internal (private) cloud
  – Cloud with in an organization

• Community cloud
  – Cloud infrastructure jointly owned by several organizations

• Public cloud
  – Cloud infrastructure owned by an organization, provided to general public as service

• Hybrid cloud
  – Composition of two or more cloud models
Short Term Implications of Clouds

- Startups and prototyping
  - Minimize infrastructure risk
  - Lower cost of entry
- Batch jobs
- One-off tasks
  - Washington post, NY Times
- Cost associatively for scientific applications
- Research at scale
Cloud Application Demand

• Many cloud applications have cyclical demand curves
  – Daily, weekly, monthly, ...

• Workload spikes are more frequent and significant
  – When some event happens like a pop star has expired:
    • More # tweets, Wikipedia traffic increases
    • 22% of tweets, 20% of Wikipedia traffic when Michael Jackson expired in 2009
      – Google thought they are under attack
Economics of Cloud Users

• Pay by use instead of provisioning for peak

Static data center

Data center in the cloud

Unused resources
Economics of Cloud Users - continued

• Risk of over-provisioning: underutilization
  – Huge sunk cost in infrastructure

Static data center
Economics of Cloud Users - continued

- Heavy penalty for under-provisioning

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Time (days)} & \text{Resources} & \text{Demand} & \text{Capacity} \\
\hline
1 & \uparrow & \uparrow & \downarrow \\
2 & \downarrow & \uparrow & \downarrow \\
3 & \downarrow & \downarrow & \uparrow \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Time (days)} & \text{Resources} & \text{Demand} & \text{Capacity} \\
\hline
1 & \uparrow & \uparrow & \downarrow \\
2 & \downarrow & \uparrow & \downarrow \\
3 & \downarrow & \downarrow & \uparrow \\
\hline
\end{array}
\]
Long Term Implications of clouds

• Application software:
  – Cloud & client parts, disconnection tolerance

• Infrastructure software:
  – Resource accounting, VM awareness

• Hardware systems:
  – Containers, energy proportionality
Cloud Computing Progress

[Armando Fox, 2010]
Our Data-driven World

- Already as of 2012, every day 2.5 exabytes ($2.5 \times 10^{18}$) of data were created [https://en.wikipedia.org/wiki/Big_data]
- Science
  - Data bases from astronomy, genomics, environmental data, Sensors and Smart Devices ...
- Humanities and Social Sciences
  - Scanned books, historical documents, social interactions data, ...
- Business & Commerce
  - Corporate sales, stock market transactions, census, airline traffic, ...
- Entertainment
  - Internet images, MP3 files, ...
- Medicine
  - MRI & CT scans, patient records, ...
- Internet of Things
  - 50 billion devices will be connected by year 2020 [Cisco]

[Agrawal et al, EDBT 2011 Tutorial]
What can we do with this wealth?

- What can we do?
  - Scientific breakthroughs
  - Business process efficiencies
  - Realistic special effects
  - Improve quality-of-life: healthcare, transportation, environmental disasters, daily life, ...

- Could We Do More?
  - YES: but need major advances in our capability to analyze this data

[Agrawal et al, EDBT 2011 Tutorial]
Platforms for Data Analysis

• Data Warehousing, Data Analytics & Decision Support Systems
• Data Analytics in the Web Context
• Data Analytics in the Cloud
Data Warehousing, Data Analytics & Decision Support Systems

- Used to manage and control business
- Transactional Data: historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and analysts to understand the business and make judgments
- Limited scalability

[Agrawal et al, EDBT 2011 Tutorial]
Information Retrieval & Data Analytics in the Web Context (Lecture 4)

- Source Selection
  - Resource
  - Query Formulation
  - Query
  - Search
  - Results
  - Selection
  - Documents
  - Examination
  - Information
  - Delivery

- System discovery
- Vocabulary discovery
- Concept discovery
- Document discovery
- Source reselection
Data Analytics in the Cloud

• Scalability to large data volumes:
  – Scan 100 TB on 1 node @ 50 MB/sec = 23 days
  – Scan on 1000-node cluster = 33 minutes

  ➔ Divide-And-Conquer (i.e., data partitioning)
    – This is where MapReduce jumps in
    – Replicate data and computation (Lectures 2, 3, 4)

• Cost-efficiency:
  – Commodity nodes (cheap, but unreliable)
  – Commodity network
  – Automatic fault-tolerance
BigData Solutions

• We mainly study Hadoop MapReduce
  – Hadoop MapReduce is one of the most used solution for large scale data analysis.
  – Spark will be discussed as an alternative (Lecture 6, 7)
    • In-Memory MapReduce

• However, Hadoop has some troubles
  – Writing low level MapReduce code is slow
  – Need a lot of expertise to optimize MapReduce code
  – Prototyping requires compilation
  – A lot of custom code is required
  – Hard to manage more complex MapReduce job chains
Pig (Lecture 5)

- Pig is a high level language on top of Hadoop MapReduce
  - Models a scripting language (Latin)
    - Fast prototyping
  - Similar to declarative SQL
    - Easier to get started
  - In comparison to Hadoop MapReduce:
    - 5% of the code
    - 5% of the time
Scaling Information Systems

• Fault tolerance, high availability & scalability are essential prerequisites for any enterprise application deployment

• Scalability
  – Generally nodes in information systems support specific load
  – When load increases beyond certain level, systems should be scaled up
  – Similarly when load decreases they should be scaled down
Scaling Information Systems - continued

• Two basic models of scaling
  – Vertical scaling
    • Also known as Scale-up
    • Achieving better performance by replacing an existing node with a much powerful machine
    • Risk of losing currently running jobs
  – Horizontal scaling
    • Aka Scale-out
    • Achieving better performance by adding more nodes to the system
    • New servers are introduced to the system to run along with the existing servers
Scaling Enterprise Applications in the Cloud

Client Site → Load Balancer (Proxy) → App Server

Client Site → Load Balancer (Proxy) → App Server

Client Site → Load Balancer (Proxy) → App Server

App Server → DB

Memcache
Load Balancer

• Load balancing has been a key mechanism in making efficient web server farms
• Load balancer automatically distributes incoming application traffic across multiple servers
• Some of the key load balancing algorithms include Random, Round Robin, Least connection
• Examples:
  – Nginx - http://nginx.org/
  – HAProxy - http://haproxy.1wt.eu/
  – Pen - http://siag.nu/pen/
Scaling in the Cloud - bottleneck

Load Balancer (Proxy)

App Server

Client Site

Database becomes the Scalability Bottleneck

Cannot leverage elasticity
Scaling in the Cloud - bottleneck

Scalable and Elastic, but limited consistency and operational flexibility

NoSQL

(Lecture 8)
This week in Lab (Homework)

• Registration to the cloud & keys
  – Working with Eucatools and API
• Some small programming task to get you started
Next lecture

• Introduction to MapReduce
References

• Several of the slides are taken from Prof. Anthony D. Joseph’s lecture at RWTH Aachen (March 2010)

• Some slides are also taken from “Big Data and Cloud Computing: Current State and Future Opportunities”, tutorial at EDBT 2011 by D. Agrawal, S. Das and A. Abbadi.