

E534 - Big Data Applications

Lecture Notes

Geoffrey C. Fox
Gregor von Laszewski

Editor

laszewski@gmail.com

<https://cloudmesh-community.github.io/book/vonlaszewski-e534.epub>

September 06, 2019 - 11:18 AM

Created by Cloudmesh & Cyberaide Bookmanager, <https://github.com/cyberaide/bookmanager>

E534 - BIG DATA APPLICATIONS

Geoffrey C. Fox Gregor von Laszewski

(c) Indiana University, Gregor von Laszewski, Geoffrey Fox, 2018, 2019

E534 - BIG DATA APPLICATIONS

1 PREFACE

1.1 Disclaimer 

1.1.1 Acknowledgment

1.1.2 Extensions

2 WEEK 1

2.1 Part I Motivation I 

2.1.1 Motivation

2.1.2 00) Mechanics of Course, Summary, and overall remarks on course

2.1.2.1 01A) Technology Hypecycle I

2.1.2.2 01B) Technology Hypecycle II

2.1.2.3 01C) Technology Hypecycle III

2.1.3 01D) Technology Hypecycle IV

2.1.4 02)

2.1.4.1 02A) Clouds/Big Data Applications I

2.1.4.2 02B) Cloud/Big Data Applications II

2.1.4.3 02C) Cloud/Big Data

2.1.5 03) Jobs In areas like Data Science, Clouds and Computer Science and Computer

2.1.6 04) Industry, Technology, Consumer Trends Basic trends 2018 Lectures 4A 4B have

2.1.7 05) Digital Disruption and Transformation The Past displaced by Digital

2.1.8 06)

2.1.9 06A) Computing Model I Industry adopted clouds which are attractive for data

2.1.9.1 06B) Computing Model II with 3 subsections is removed; please see 2018

2.1.10 07) Research Model 4th Paradigm; From Theory to Data driven science?

2.1.11 08) Data Science Pipeline DIKW: Data, Information, Knowledge, Wisdom, Decisions.

2.1.12 09) Physics: Looking for Higgs Particle with Large Hadron Collider LHC Physics as a big data example

2.1.13 10) Recommender Systems I General remarks and Netflix

example

2.1.14 11) Recommender Systems II Exploring Data Bags and Spaces

2.1.15 12) Web Search and Information Retrieval Another Big Data Example

2.1.16 13) Cloud Applications in Research Removed Science Clouds, Internet of Things

2.1.17 14) Parallel Computing and MapReduce Software Ecosystems

2.1.18 15) Online education and data science education Removed.

2.1.19 16) Conclusions

3 WEEK 2

3.1 Part II Motivation II ☁

3.1.1 2018 BDAA Motivation-1A) Technology Hypecycle I

3.1.2 2018 BDAA Motivation-1B) Technology Hypecycle II

3.1.3 2018 BDAA Motivation-2B) Cloud/Big Data Applications II

3.1.4 2018 BDAA Motivation-4A) Industry Trends I

3.1.5 2018 BDAA Motivation-4B) Industry Trends II

3.1.6 2017 BDAA Motivation-4C) Industry Trends III

3.1.7 2018 BDAA Motivation-6B) Computing Model II

3.1.8 2017 BDAA Motivation-8) Data Science Pipeline DIKW

3.1.9 2017 BDAA Motivation-13) Cloud Applications in Research Science Clouds Internet of Things

3.1.10 2017 BDAA Motivation-15) Data Science Education Opportunities at Universities

4 WEEK 3

4.1 Part III Cloud ☁

4.1.1 A. Summary of Course

4.1.2 B. Defining Clouds I

4.1.3 C. Defining Clouds II

4.1.4 D. Defining Clouds III: Cloud Market Share

4.1.5 E. Virtualization: Virtualization Technologies,

4.1.6 F. Cloud Infrastructure I

4.1.7 G. Cloud Infrastructure II

4.1.8 H. Cloud Software:

4.1.9 I. Cloud Applications I: Clouds in science where area called

4.1.10 J. Cloud Applications II: Characterize Applications using NIST

4.1.11 K. Parallel Computing

4.1.12 L. Real Parallel Computing: Single Program/Instruction Multiple

Data SIMD SPMD

4.1.13 M. Storage: Cloud data

4.1.14 N. HPC and Clouds

4.1.15 O. Comparison of Data Analytics with Simulation:

4.1.16 P. The Future I

4.1.17 Q. other Issues II

4.1.18 R. The Future and other Issues III

5 REFERENCES

1 PREFACE

Fri Sep 6 11:18:05 EDT 2019 

1.1 DISCLAIMER

This book has been generated with [Cyberaide Bookmanager](#).

Bookmanager is a tool to create a publication from a number of sources on the internet. It is especially useful to create customized books, lecture notes, or handouts. Content is best integrated in markdown format as it is very fast to produce the output.

Bookmanager has been developed based on our experience over the last 3 years with a more sophisticated approach. Bookmanager takes the lessons from this approach and distributes a tool that can easily be used by others.

The following shields provide some information about it. Feel free to click on them.

1.1.1 Acknowledgment

If you use bookmanager to produce a document you must include the following acknowledgement.

“This document was produced with Cyberaide Bookmanager developed by Gregor von Laszewski available at <https://pypi.python.org/pypi/cyberaide-bookmanager>. It is in the responsibility of the user to make sure an author acknowledgement section is included in your document. Copyright verification of content included in a book is responsibility of the book editor.”

The bibtex entry is

```
@Misc{www-cyberaide-bookmanager,
```

```
author =  {Gregor von Laszewski},  
title =   {{Cyberaide Book Manager}},  
howpublished = {pypi},  
month =    apr,  
year =     2019,  
url={https://pypi.org/project/cyberaide-bookmanager/}  
}
```

1.1.2 Extensions

We are happy to discuss with you bugs, issues and ideas for enhancements.
Please use the convenient github issues at

- <https://github.com/cyberaide/bookmanager/issues>

Please do not file with us issues that relate to an editors book. They will provide you with their own mechanism on how to correct their content.

2 WEEK 1

2.1 PART I MOTIVATION I



2.1.1 Motivation

Big Data Applications & Analytics: Motivation/Overview; Machine (actually Deep) Learning, Big Data, and the Cloud; Centerpieces of the Current and Future Economy,

2.1.2 00) Mechanics of Course, Summary, and overall remarks on course

In this section we discuss the summary of the motivation section.

The screenshot shows a presentation slide with a black header and footer. The main content area contains several bullet points and a small video thumbnail at the bottom right. The bullet points include:

- **Technology Hypecycles**: especially that for emerging technologies in 2019
- Details of 2019 Emerging Technology and related 4th Gartner Hypecycle
- Details of 2019 Gartner Hypecycles
- Gartner Hypecycles and Priority Matrices for emerging technologies in 2018, 2017 and 2016
- More details on 2018 will be found in Unit 1A of 2018 Presentations and details of 2017 Unit 1B and 2016 Unit 1B will be found in later years
- **Gartner Technology Hypecycles IV**
- Details of 2018 Hypecycles and Priority matrix at selected time in 2018-2019
- Details of 2019 Hypecycles and Priority matrix at selected time in 2019-2020
- They are mixed up with transformational and disruptive changes
- Only the first few pages have more details of this history including Priority matrices

2.1.2.1 01A) Technology Hypecycle I

Today clouds and big data have got through the hype cycle (they have emerged) but features like blockchain, serverless and machine learning are on recent hype cycles while areas like deep learning have several entries (as in fact do clouds) Gartner's Hypecycles and especially that for emerging technologies in 2019 The phases of hypecycles Priority Matrix with benefits and adoption time Initial discussion of 2019 Hypecycle for Emerging Technologies

The screenshot shows a presentation slide with a black header and footer. The main content area contains two diagrams: 'Hype Curves of the Hype Cycle' and 'Priority Matrix'. The 'Hype Curves' diagram shows a bell-shaped curve peaking at 'Peak of Inflated Expectations'. The 'Priority Matrix' diagram is a grid with columns labeled 'Priority Matrix' and 'Value or maximum category'. It includes a legend for colors: red (High), orange (Medium-High), yellow (Medium-Low), green (Medium-High), and blue (Low). A small video thumbnail is at the bottom left.

2.1.2.2 01B) Technology Hypecycle II

Today clouds and big data have got through the hype cycle (they have emerged)

but features like blockchain, serverless and machine learning are on recent hype cycles while areas like deep learning have several entries (as in fact do clouds) Gartner's Hypecycles and especially that for emerging technologies in 2019 Details of 2019 Emerging Technology and related (AI, Cloud) Hypecycles



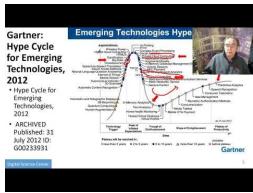
2.1.2.3 01C) Technology Hypecycle III

Today clouds and big data have got through the hype cycle (they have emerged) but features like blockchain, serverless and machine learning are on recent hype cycles while areas like deep learning have several entries (as in fact do clouds) Gartners Hypecycles and Priority Matrices for emerging technologies in 2018, 2017 and 2016 More details on 2018 will be found in Unit 1A of 2018 Presentation and details of 2015 in Unit 1B (Journey to Digital Business). 1A in 2018 also discusses 2017 Data Center Infrastructure removed as this hype cycle disappeared in later years.



2.1.3 01D) Technology Hypecycle IV

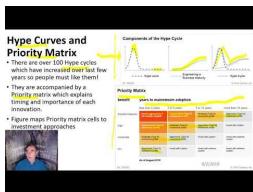
Today clouds and big data have got through the hype cycle (they have emerged) but features like blockchain, serverless and machine learning are on recent hype cycles while areas like deep learning have several entries (as in fact do clouds) Emerging Technologies hypecycles and Priority matrix at selected times 2008-2015 Clouds star from 2008 to today They are mixed up with transformational and disruptive changes Unit 1B of 2018 Presentation has more details of this history including Priority matrices



2.1.4 02)

2.1.4.1 02A) Clouds/Big Data Applications I

The Data Deluge Big Data; a lot of the best examples have NOT been updated (as I can't find updates) so some slides old but still make the correct points Big Data Deluge has become the Deep Learning Deluge Big Data is an agreed fact; Deep Learning still evolving fast but has stream of successes!



2.1.4.2 02B) Cloud/Big Data Applications II

Clouds in science where area called cyberinfrastructure; The usage pattern from NIST is removed. See 2018 lectures 2B of the motivation for this discussion



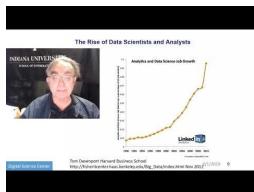
2.1.4.3 02C) Cloud/Big Data

Usage Trends Google and related Trends Artificial Intelligence from Microsoft, Gartner and Meeker



2.1.5 03) Jobs In areas like Data Science, Clouds and Computer Science and Computer

Engineering



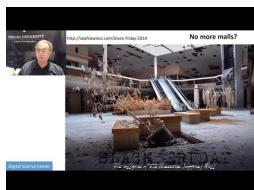
2.1.6 04) Industry, Technology, Consumer Trends Basic trends 2018 Lectures 4A 4B have

more details removed as dated but still valid See 2018 Lesson 4C for 3 Technology trends for 2016: Voice as HCI, Cars, Deep Learning



2.1.7 05) Digital Disruption and Transformation The Past displaced by Digital

Disruption; some more details are in 2018 Presentation Lesson 5



2.1.8 06)

2.1.9 06A) Computing Model I Industry adopted clouds which are attractive for data

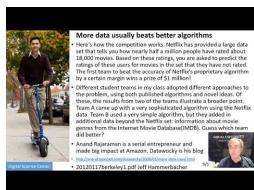
analytics. Clouds are a dominant force in Industry. Examples are given

2.1.9.1 06B) Computing Model II with 3 subsections is removed; please see 2018

Presentation for this Developments after 2014 mainly from Gartner Cloud Market share Blockchain

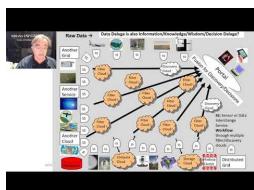


2.1.10 07) Research Model 4th Paradigm; From Theory to Data driven science?



2.1.11 08) Data Science Pipeline DIKW: Data, Information, Knowledge, Wisdom, Decisions.

More details on Data Science Platforms are in 2018 Lesson 8 presentation



2.1.12 09) Physics: Looking for Higgs Particle with Large Hadron Collider LHC Physics as a big data example



2.1.13 10) Recommender Systems I General remarks and Netflix example



2.1.14 11) Recommender Systems II Exploring Data Bags and Spaces



2.1.15 12) Web Search and Information Retrieval Another Big Data Example

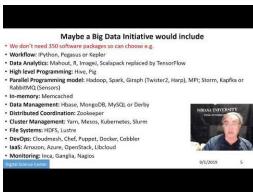


2.1.16 13) Cloud Applications in Research Removed Science Clouds, Internet of Things

Part 12 continuation. See 2018 Presentation (same as 2017 for lesson 13) and Cloud Unit 2019-I) this year



2.1.17 14) Parallel Computing and MapReduce Software Ecosystems



2.1.18 15) Online education and data science education Removed.

You can find it in the 2017 version. In [Week 2](#) you can see more about this.



2.1.19 16) Conclusions

Conclusion contain in the latter part of the part 15.

Motivation Archive Big Data Applications & Analytics: Motivation/Overview; Machine (actually Deep) Learning, Big Data, and the Cloud; Centerpieces of the Current and Future Economy. Backup Lectures from previous years referenced in 2019 class

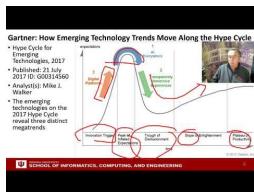


3 WEEK 2

3.1 PART II MOTIVATION II

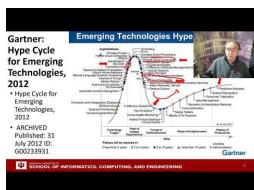
3.1.1 2018 BDAA Motivation-1A) Technology Hypecycle I

In this section we discuss on general remarks including Hype curves.



3.1.2 2018 BDAA Motivation-1B) Technology Hypecycle II

In this section we continue our discussion on general remarks including Hype curves.



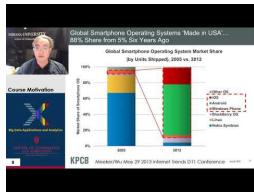
3.1.3 2018 BDAA Motivation-2B) Cloud/Big Data Applications II

In this section we discuss clouds in science where area called cyberinfrastructure; the usage pattern from NIST Artificial Intelligence from Gartner and Meeker.



3.1.4 2018 BDAA Motivation-4A) Industry Trends I

In this section we discuss on Lesson 4A many technology trends through end of 2014.



3.1.5 2018 BDAA Motivation-4B) Industry Trends II

In this section we continue our discussion on industry trends. This section includes Lesson 4B 2015 onwards many technology adoption trends.



3.1.6 2017 BDAA Motivation-4C) Industry Trends III

In this section we continue our discussion on industry trends. This section contains lesson 4C 2015 onwards 3 technology trends voice as HCI cars deep learning.



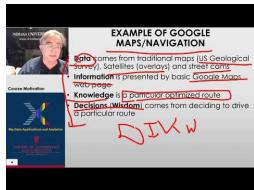
3.1.7 2018 BDAA Motivation-6B) Computing Model II

In this section we discuss computing models. This section contains lesson 6B with 3 subsections developments after 2014 mainly from Gartner cloud market share blockchain



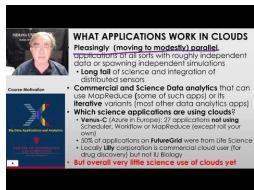
3.1.8 2017 BDAA Motivation-8) Data Science Pipeline DIKW

In this section, we discuss data science pipelines. This section also contains about data, information, knowledge, wisdom forming DIKW term. And also it contains some discussion on data science platforms.



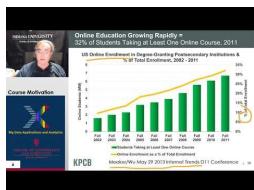
3.1.9 2017 BDAA Motivation-13) Cloud Applications in Research Science Clouds Internet of Things

In this section we discuss about internet of things and related cloud applications.



3.1.10 2017 BDAA Motivation-15) Data Science Education Opportunities at Universities

In this section we discuss more on data science education opportunities.



4 WEEK 3

4.1 PART III CLOUD

4.1.1 A. Summary of Course

Overall Summary II

- I) Cloud Infrastructure: Comments on trends in the data center and its technologies
 - Cloud physically across the world
 - Cloud computing
 - Fraction of world's computing resources in clouds and associated costs
 - An analysis from Gartner and other consulting firms
- G) Cloud Infrastructure: Gartner forecast and priority matrix on Compute Infrastructure
 - Containers composed to virtual machine
 - The emergence of containerized computing as a dominant form
- H) Cloud Software: (PaaS) with over 300 software packages and how to use each of 21 layers
 - Google's software innovations
 - Machine learning
 - Cloud and HPC software stacks composed
 - Components need to support cloud/distributed system programming

Watch Lecture



4.1.2 B. Defining Clouds I

In this lecture we discuss the basic definition of cloud and two very simple examples of why virtualization is important.

Overall Summary II

- I) Cloud Infrastructure: Comments on trends in the data center and its technologies
 - Cloud physically across the world
 - Cloud computing
 - Fraction of world's computing resources in clouds and associated costs
 - An analysis from Gartner and other consulting firms
- G) Cloud Infrastructure: Gartner forecast and priority matrix on Compute Infrastructure
 - Containers composed to virtual machine
 - The emergence of containerized computing as a dominant form
- H) Cloud Software: (PaaS) with over 300 software packages and how to use each of 21 layers
 - Google's software innovations
 - Machine learning
 - Cloud and HPC software stacks composed
 - Components need to support cloud/distributed system programming

Watch Lecture



In this lecture we discuss how clouds are situated wrt HPC and supercomputers, why multicore chips are important in a typical data center.

4.1.3 C. Defining Clouds II

In this lecture we discuss service-oriented architectures, Software services as Message-linked computing capabilities.

Different aaS (as aService)'s

- IaaS: Infrastructure is "renting" service for hosts
- PaaS: Convenient service interface to develop applications
- SaaS: Convenient service interface to usage
 - New "Compute as a Service" fault applies at application level
- NaaS: Summarizes modern "Software Defined" services
- Support Computing as a service is "my infrastructure in the cloud"

Support Computing

- Support Computing
- Support Computing
- Support Computing

SaaS

- Qualified Images
- Creating Images
- Actual Images

Refactoring

- Open Source Applications

Platform

- Cloud e.g. AWS Lambda

PaaS

- HTTP e.g. REST, SOAP

Infra

- Software Defined
- Compute, Storage, Network, Hypervisor, Base Metal

IaaS

- Infrastructure as a Service

NaaS

- Network as a Service

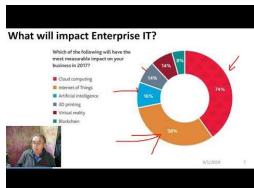
Watch Lecture



In this lecture we discuss different aaS's: Network, Infrastructure, Platform, Software. The amazing services that Amazon AWS and Microsoft Azure have Initial Gartner comments on clouds (they are now the norm) and evolution of servers; serverless and microservices Gartner hypecycle and priority matrix on

Infrastructure Strategies.

4.1.4 D. Defining Clouds III: Cloud Market Share



In this lecture we discuss on how important the cloud market shares are and how much money do they make.

4.1.5 E. Virtualization: Virtualization Technologies,



In this lecture we discuss hypervisors and the different approaches KVM, Xen, Docker and Openstack.

4.1.6 F. Cloud Infrastructure I



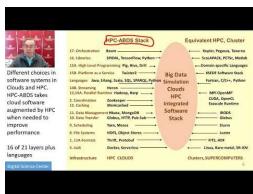
In this lecture we comment on trends in the data center and its technologies. Clouds physically spread across the world Green computing Fraction of world's computing ecosystem. In clouds and associated sizes an analysis from Cisco of size of cloud computing is discussed in this lecture.

4.1.7 G. Cloud Infrastructure II



In this lecture, we discuss Gartner hypecycle and priority matrix on Compute Infrastructure Containers compared to virtual machines The emergence of artificial intelligence as a dominant force.

4.1.8 H. Cloud Software:



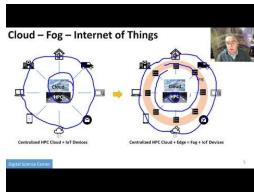
In this lecture we discuss, HPC-ABDS with over 350 software packages and how to use each of 21 layers Google's software innovations MapReduce in pictures Cloud and HPC software stacks compared Components need to support cloud/distributed system programming.

4.1.9 I. Cloud Applications I: Clouds in science where area called



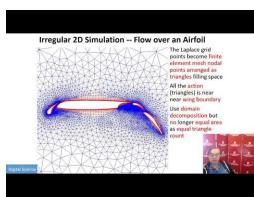
In this lecture we discuss cyberinfrastructure; the science usage pattern from NIST Artificial Intelligence from Gartner.

4.1.10 J. Cloud Applications II: Characterize Applications using NIST



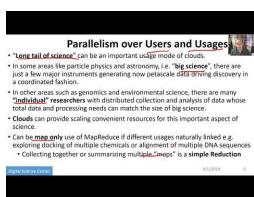
In this lecture we discuss the approach Internet of Things with different types of MapReduce.

4.1.11 K. Parallel Computing



In this lecture we discuss analogies, parallel computing in pictures and some useful analogies and principles.

4.1.12 L. Real Parallel Computing: Single Program/Instruction Multiple Data SIMD SPMD



In this lecture, we discuss Big Data and Simulations compared and we furthermore discusses what is hard to do.

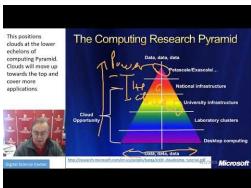
4.1.13 M. Storage: Cloud data



In this lecture we discuss about the approaches, repositories, file systems, data

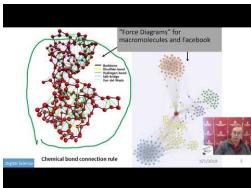
lakes.

4.1.14 N. HPC and Clouds



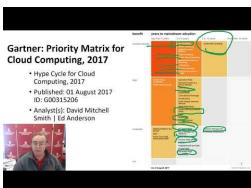
In this lecture we discuss the Branscomb Pyramid Supercomputers versus clouds Science Computing Environments.

4.1.15 O. Comparison of Data Analytics with Simulation:



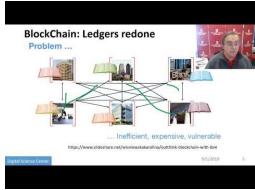
In this lecture we discuss the structure of different applications for simulations and Big Data Software implications Languages.

4.1.16 P. The Future I



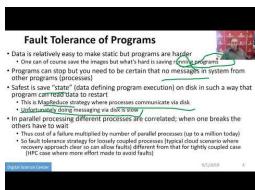
In this lecture we discuss Gartner cloud computing hypecycle and priority matrix 2017 and 2019 Hyperscale computing Serverless and FaaS Cloud Native Microservices Update to 2019 Hypecycle.

4.1.17 Q. other Issues II



In this lecture we discuss on Security Blockchain.

4.1.18 R. The Future and other Issues III



In this lecture we discuss on Fault Tolerance.

5 REFERENCES

