CS33001: DATA-INTENSIVE COMPUTING SYSTEMS SEMINAR

Today

• Discuss Papers
• Discuss platform / infrastructure choices
• Discuss infrastructure assessment assignment

April 5, 2013
CS33001 Chien Spring 2013

READINGS

For Friday, 4/4
  • http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf


For Monday, 4/9 (Erik Bodzsar, Guest lecture)
Presto https://sites.google.com/site/uchicagolssg/lssg/research/blockus

For Friday, 4/12, Data-Parallel archetypes
  • Map Reduce http://research.google.com/archive/mapreduce.html

For Monday, 4/15 (Andrew Baptist, Guest Lecture)
Cleversafe (tbd)

April 5, 2013
CS33001 Chien Spring 2013
DIGITAL UNIVERSE DISCUSSION

Assessment of the size of “data” in the world, data is growing rapidly
2011: 1.8 trillion gigabytes = \(10^{12} \times 10^9\) => \(10^{21}\) (zettabytes = \(2^{70}\))... Doubling slightly faster than every 2 years
⇒ Storage is important (keep it)
⇒ Extract information from it
⇒ Tag it, organize it
⇒ Keep others from seeing it (security)
Generation increasing faster than storage density
Methodology is problematic

April 5, 2013
CS33001 Chien Spring 2013

SOURCES OF DATA

70% of data is from personal activity; corporations held 80% of this data
⇒ pictures on phones and uploading, Flickr, facebook/instagram, gmail, => can be analyzed in aggregate
⇒ video (personal, commercial), image (the largest volume)
⇒ electronic activities (sensors/rfid, transactional, organizational, open continuous sensors, everything you do)
⇒ web tracking... Other network activity tracking
⇒ medical
⇒ consumer activity (primary and digital shadow)
⇒ also networked devices drive increase in data

April 5, 2013
CS33001 Chien Spring 2013
WHERE IS THE DATA?

Claim 10% is maintained in the cloud, 20% is touched by the cloud – claim: growing to 33% by 2020
IT staff not growing relative to amount of data
- Each IT staff member must manage more data each year
- Pitching at CIO’s who have limited staff
Not all data is equally important
- LHC, enterprise, frivolous pictures
Massive Duplication
- In cloud services, can exploit this for efficiency
- Sharing copies for non-unique data
- Privacy issues in sharing (encrypted data)
- Fingerprinting to avoid network bandwidth

3 GOOD

Great ambition, how much data is there? (and how fast is it growing)
Assessment of where the data is coming from (devices, personal activity, image/video)
Much of data is not unique (economies of aggregation)
- disincentive for encryption
Growth storage technology/sales/markets
Who has the data personal/enterprise/cloud (as overlapping categories)
Digital shadow (data you might not be aware exists) – and all of the followon processing
3 BAD

No clear definition of what data is? (anything you could extract information from; every bit ever generated; persisted data – files, database, etc.)
Actively misleading figures for commercial purpose (form of presentation)
Gap between data creation and capacity (sales, products)
Blatant commercialism
General lack of rigor (apparent lack of rigor)

DATA DWARVES
DISCUSSION

An attempt to characterize and categorize a data-centric workloads. They are different from “others”, they are different from TPC benchmarks. And create a set of “dwarves” to support systems research
- Key dimension of attributes
- Fundamental (7 dwarves), or simply covering... Its not clear if they compose...
- sort, search, dedup, recommend, video sharing, (data mining, transaction)

Can we use for systems/architecture evaluation?
Can we derive canonical modules or capabilities that can be composed into these Data-intensive applications?
3 GOOD AND BAD

Great idea – find canonical problems
Brought some breadth; abstracted from a particular computer system architecture, or expression
Reasonably up to date (at time or writing)

Didn't completely succeed
Don’t explain how to choose a coverage set
No realization of the dwarves

PROJECT ASSIGNMENT

Identify a challenging data-intensive computing project and read up on it

• What defines it as a data-intensive computing project? (as opposed to something-else intensive)
• What are some of the unique technical challenges it represents? Systems challenges?
• What is the value of having all that data? Summaries? (there’s clearly a cost)
• What are some unique opportunities it represents? Where do the timeliness/quality/yield requirements come from?
• If significant improvements were possible? (speed/quality/cost) What if any new opportunities would it unlock?
• What computing infrastructure are they using? Is it efficient? Is it accessible?
CANDIDATES

HEP Data – ATLAS
Montage, EOSDIS Earth-observation system (NASA)
Glass Phase
1000Genomes – Phylogeny
Metagenomic Assembly ( ) => KBASE
Andrei Rhzetsky’s work
Netflix – recommender systems for movies
Consumer credit card fraud detection (public services? – social services chapin hall)
GWAS (Genome wide Association) – genome based medicine
OpenMap
Chicago Open Data project – public governance transparency
Facebook (to make better advertising)
Traffic real-time
Government/DHS finding adversaries

ASSIGNMENT FORMAT
(4/12, TURN-IN NIGHT OF 4/11)

3-page writeup describing a data-intensive computing project and its goals (and answer the list of questions)
=> Typically a large-scale ambitious effort

We’ll select a subset of the most interesting writeups to discuss in class

• General information
• Status, impact on application/science/commerce
• Impact on systems
• Can it be leveraged into a course project
PROJECT ASSIGNMENT
(MONDAY 4/15)

Download, install, and run a data-intensive computing infrastructure

• A widely used one? (MongoDB, Hbase/H*, Graphlab, Cassandra)
• Or get started with Presto/Blockus or Cleversafe
• What is it capable of?
• What types of problems is it particularly well suited to? Intended workload?
• Does it scales? (in data? In speed/capability?) does it scale down?
• Robustness/Resilience of the system – hw/sw, operating point/ usage, does it degrade or collapse?
• Recovery and Diagnosis – what can you recover in a failure? And what can you deduce about the cause of the failure?
• What kind of hardware was designed for? (clusters, HPC) – communication, reliability, system balance issues. Distribution?
• Is it efficient? (cost, energy, algorithmically, human effort)

CANDIDATES

HBASE/H*, VoltDB
PIG/H*
HadoopDB/H*
Cassandra
SciDB
BLOOM/MR Online/?
MongoDB
Graphlab/Graphchi
Swift
?

Preference: something new
ASSIGNMENT TURN-IN FORMAT (4/15)

1-page writeup describing system and its capabilities
5-minute presentation in class using 4 slides – summarize capabilities and your experience with it (what you did)

• For each, we’ll have a discussion on what it’s being used for
• What it’s good at
• What are its shortcomings
• What kinds of projects it might be suitable for

SUMMARY

Data Universe
Data Dwarfs
Reminders
- Data intensive computing infrastructure Assignment
- Readings (Presto paper)
- Guest lecture: Erik Bodzsar
GROUND RULES FOR THE COURSE

No “tourists” – come and come regularly
Active participation – come prepared, and come with something to say, and with questions to be answered
Push the envelope – beyond the questions framed in the papers, ideas in projects, to their logical extreme or conclusion
No “sacred cows” – any and all technical (and even ecosystem) topics can be opened and discussed (Andrew’s call to shape discussion based on “productivity”)