Introduction to Big Data Management and UDBMS research in Helsinki

Jiaheng Lu
Department of Computer Science
University of Helsinki
Big number, small number – from data to understanding

DATA
IS THE NEW OIL
but do you have the resource to refine it?
Outline

• Introduction to Big Data

• Cloud computing

• MapReduce programming model

• Our research on multi-model databases and big data
Four V’s

- Volume: Data Size
- Velocity: Speed of Change
- Variety: Different Forms of Data Sources
- Veracity: Uncertainty of Data

Big Data
Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 to 2020
  - From 0.8 Zettabytes to 35 Zb
  - Data volume is increasing exponentially
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …

To extract knowledge ➔ all these types of data need to linked together
Velocity (Speed)

- Data is generated fast and needs to be processed fast
- Late decisions ➔ missing opportunities

**Examples**

- **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you

- **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction
Big data 4V’s

**Volume**
- Data at Rest
  - Terabytes to exabytes of existing data to process

**Velocity**
- Data in Motion
  - Streaming data, milliseconds to seconds to respond

**Variety**
- Data in Many Forms
  - Structured, unstructured, text, multimedia

**Veracity**
- Data in Doubt
  - Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
Big data technologies

Data Management

Compute

Storage

Distributions & Data Warehouse

Data Management should abstract developers from technology change
Two technologies on Big Data Management

- Cloud computing
- Hadoop and MapReduce
Why we use cloud computing?
Why we use cloud computing?

Case 1:  
Write a file  
Save  
Computer down, file is lost  

Files are always stored in cloud, never lost
Why we use cloud computing?

Case 2:
Use MS Word --- download, install, use
Use Skype --- download, install, use
Use C++ IDE --- download, install, use
……

Get the serve from the cloud
What is cloud and cloud computing?

Cloud
 Demand resources or services over Internet scale and reliability of a data center.
What is cloud and cloud computing?

Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet.
Characteristics of cloud computing

- **Virtual.**
  software, databases, Web servers, operating systems, storage and networking as virtual servers.

- **On demand.**
  add and subtract processors, memory, network bandwidth, storage.
Types of cloud service

- **SaaS**
  Software as a Service

- **PaaS**
  Platform as a Service

- **IaaS**
  Infrastructure as a Service
- **SaaS**
  - Software as a service
  - Operating environment largely irrelevant, fully functional applications provided, e.g. CRM, ERP, email

- **PaaS**
  - Platform as a service
  - Operating environment included, e.g. Windows/.NET, Linux/J2EE, applications of choice deployed

- **IaaS**
  - Infrastructure as a service
  - Virtual platform on which required operating environment and application are deployed
  - Includes storage as a service offerings
Two technologies on Big Data Management

- Cloud computing
- Hadoop and MapReduce
What is Hadoop?

• Apache top level project, open-source implementation of frameworks for reliable, scalable, distributed computing and data storage.
Google Origins

The Google File System
Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google

MapReduce: Simplified Data Processing on Large Clusters
Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.

Bigtable: A Distributed Storage System for Structured Data
Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows,-Tushar Chandra, Andrew Fikes, Robert E. Gruber
info@cs.berkeley.edu, http://research.google.com
Google, Inc.

Abstract
Bigtable is a distributed storage system for managing terabyte of data that is designed to scale to a very large production of data across hundreds of commodity servers. Many projects at Google store data in Bigtable, including Google Earth and Google Fi

2003
2004
2006
Hadoop’s Developers

2005: Doug Cutting and Michael J. Cafarella developed Hadoop to support distribution for the Nutch search engine project.

The project was funded by Yahoo.

2006: Yahoo gave the project to Apache Software Foundation.
Some Hadoop Milestones

- **2008 - Hadoop Wins Terabyte Sort Benchmark** (sorted 1 terabyte of data in 209 seconds, compared to previous record of 297 seconds)

- **2010 - Hadoop's Hbase, Hive and Pig subprojects completed**, adding more computational power to Hadoop framework

- **2013 - Hadoop 1.1.2 and Hadoop 2.0.3 alpha.**
  - Ambari, Cassandra, Mahout have been added

- **2016 - Hadoop 3.0.0 Alpha-1**
Introduction to MapReduce
MapReduce: Insight

- "Consider the problem of counting the number of frequency of each word in a large collection of documents"

- Word-count problem
Simple example: Word count

Each mapper receives some of documents as input.
Simple example: Word count

Each mapper receives some of documents as input

Mappers process the KV-pairs.
Simple example: Word count

1. Each mapper receives some of documents as input
2. Mappers process the KV-pairs.
3. Each KV-pair output by the mapper is sent to the reducer

( Finland) ( Finland, 1)
( Sweden) ( Finland, 1)
( Finland) ( Germany, 1)
( Denmark, 1)

(Mapper (1-2)) (Reducer (A-G))

(Norway) ( Norway, 1)
(Germany) ( Norway, 1)

(Mapper (3-4)) (Reducer (H-N))

(Russia) ( Sweden, 1)
(Denmark) ( Sweden, 1)
(Ukraine) ( Sweden, 1)

(Mapper (5-6)) (Reducer (O-U))

(Norway, 1)
(Denmark, 1)
( Sweden, 1)
(Russia, 1)
(Sweden, 1)
(Ukraine, 1)
Simple example: Word count

1. Each mapper receives some of documents as input
2. Mappers process the KV-pairs.
3. Each KV-pair output by the mapper is sent to the reducer
4. The reducers sort their input by key

Mapper (1-2) -> Reducer (A-G)
Mapper (3-4) -> Reducer (H-N)
Mapper (5-6) -> Reducer (O-U)

(Denmark, 1)
(Finland, 1)
(Germany, 1)
(Norway, 1)
(Russia, 1)
(Sweden, 1)
(Ukraine, 1)

(Finland)
(Sweden Finland)
(Norway Germany)
(Russia Denmark)
(Sweden Ukraine)
Simple example: Word count

1. Each mapper receives some of documents as input
2. Mappers process the KV-pairs.
3. Each KV-pair output by the mapper is sent to the reducer
4. The reducers sort their input by key
5. The reducers process their input
MapReduce dataflow

The intermediate (key, value) pairs are processed by a series of mappers and reducers, with the intermediate data being shuffled and then reduced to produce the final output.

"The Shuffle"
map(String input_key, String input_value):
   // input_key: document name
   // input_value: document contents
   for each word w in input_value:
       EmitIntermediate(w, "1");

reduce(String output_key, Iterator intermediate_values):
   // output_key: a word
   // output_values: a list of counts
   int result = 0;
   for each v in intermediate_values:
       result += parseInt(v);
   Emit(AsString(result));
MapReduce: Example

Input

Intermediate

Group by Key

Grouped

Output
Outline

• Introduction to Big Data
• Cloud computing
• MapReduce programming model
• Our research on multi-model databases
A grand challenge on Variety

- Big data: Volume, Variety, Velocity, Veracity
- Variety: tree data (XML, JSON), graph data (RDF, property graphs, networks), tabular data (CSV), temporal and spatial data, text

NoSQL database types

<table>
<thead>
<tr>
<th>Types of NoSQL DBs</th>
<th>Neo4j</th>
<th>Titan</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graph Database</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Key Value Database</strong></td>
<td>Amazon DynamoDB</td>
<td>Cassandra</td>
</tr>
<tr>
<td><strong>Column Database</strong></td>
<td>Apache HBase</td>
<td>Google BigTable</td>
</tr>
<tr>
<td><strong>Document Database</strong></td>
<td>CouchDB</td>
<td>MongoDB</td>
</tr>
</tbody>
</table>

Multi-model DB

- One unified database for multi-model data
Multi-model databases

• A multi-model database is designed to support multiple data models against a single, integrated backend.

• Document, graph, relational, and key-value models are examples of data models that may be supported by a multi-model database.
Conclusion

Big data era: Volume, Variety, Velocity, Veracity

Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet.

MapReduce is a software programming model for distributed big data processing
Task on data analysis for computer linguistic model

A data processing task for computational linguistic model.

Each group will be given an article, and the students need to complete the following three steps to visualize and analyze the document.
References

References

References

(9) Jiaheng Lu, Chunbin Lin, Wei Wang, Chen Li, Haiyong Wang: String similarity measures and joins with synonyms. SIGMOD Conference 2013: 373-384


(11) Jiaheng Lu, Ying Lu, Gao Cong: Reverse spatial and textual k nearest neighbor search. SIGMOD Conference 2011: 349-360

(12) Jiaheng Lu, Jialong Han, Xiaofeng Meng: Efficient algorithms for approximate member extraction using signature-based inverted lists. CIKM 2009: 315-324
Matemaattis-luonnontieteellinen tiedekunta / Henkilön nimi / Esityksen nimi