POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Week 1-1

1. Introduction to Big data

Big Data

Prof. Hwanjo Yu
POSTECH

Big data in real world

- Big data in the movies

- Big data in the sports

- Big data in the hospitals

Q: What would be volume and financial impact be if we were to hire another cardiovascular surgeon?

Q: What are re-admission patterns for heart failure patients?

Q: For a specific diagnosis, what are core interventions that improves the outcomes?

Big data in real world

- Government
  - “Pillbox” project in US -> reduce expenses of 50 million USD per year
  - Customized employment using Big data in Germany -> reduce 10 billion euro per 3 years
  - Open competition by NIH -> detect geographical epidemic diseases via twitter analysis
- Industry
  - Google: predict geographical epidemic flu (trajectory) via search engine log analysis
  - Google: real time road traffic service
  - Volvo: find initial faulty of newly released vehicles via SNS and blog analysis (prevent recall of 50 thousand vehicles)
  - Hertz: review customers’ evaluations by Big data analysis
  - Posco: determine the purchase time and price of raw materials
  - Watcha: movie recommendation via taste analysis (no 1., larger than Naver movie)
  - Xerox: recruiting via SNS analysis
- Prompt response to commercial condition changes, improve credibility and image, reduce expenses, improve productivity, facilitate administration, …

Big data?

- Extremely large data (Wikipedia, Mckinsey)
  - Too large to store, manage, and analyze in existing ways using existing storage and existing DBMS SWs
- Government and Industry
  - Information technology to predict trends and respond proactively.
  - Technology to collect, store, manage, search, and analyze large scale data.

Evolution of science

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a theoretical component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, computational branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes.

Jim Gray and Alex Szalay, The World Wide Telescope: An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002

Big data history

Relational database management

Data warehousing

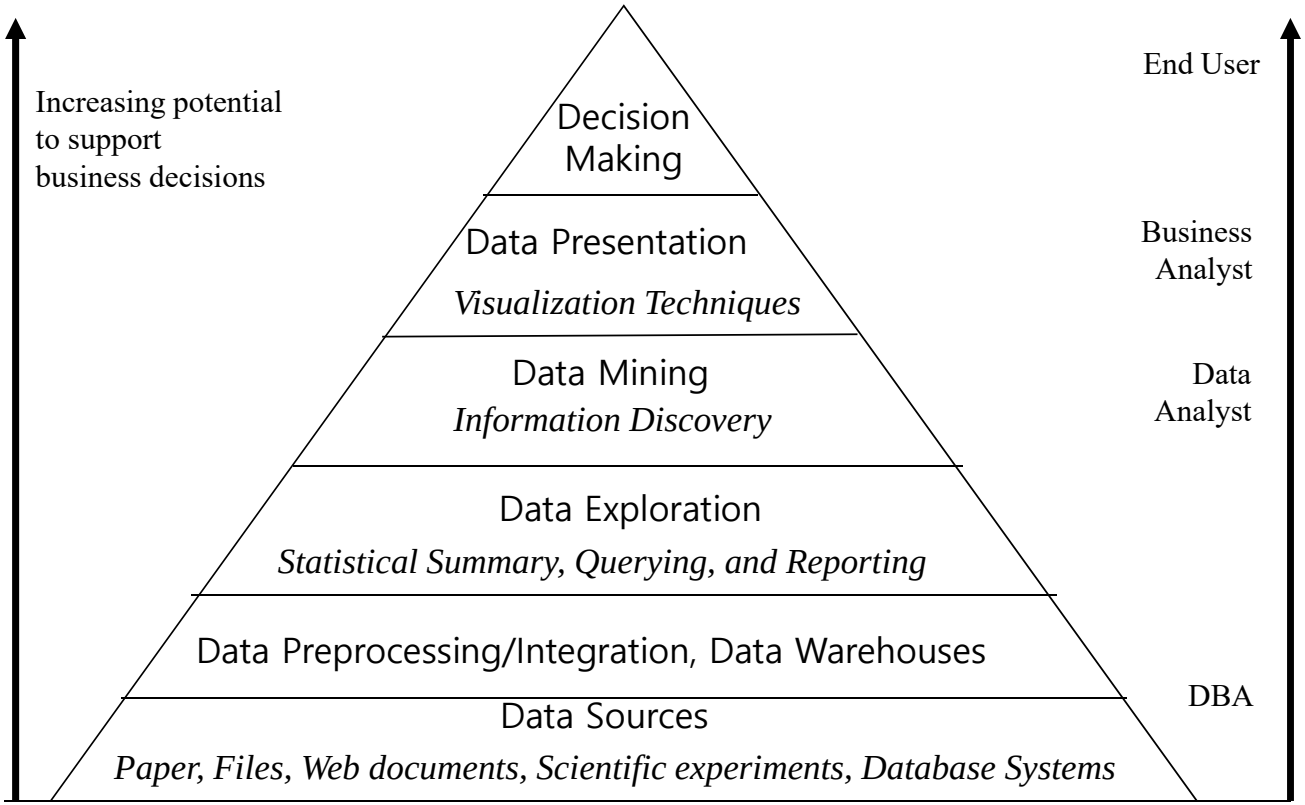
Brown, from BridgeGate LLC said,
"Many companies have implemented a data warehouse...starting to look at what they can do with all that data"

Data mining

Internet, Mobile computing, SNS, …

Big data

Business intelligence



All sciences are data sciences!

…

2.Michel J-P, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) ***Quantitative analysis of culture using millions of digitized books***. **Science** 331: 176–182. doi: 10.1126/science.1199644. Find this article online

3.Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007) ***Quantifying the evolutionary dynamics of language***. **Nature** 449: 713– 716. doi: 10.1038/nature06137. Find this article online

4.Pagel M, Atkinson QD, Meade A (2007) ***Frequency of word-use predicts rates of lexical evolution throughout Indo-European history***. **Nature** 449: 717–720. doi: 10.1038/nature06176. Find this article online

…

6. DeWall CN, Pond RS Jr, Campbell WK, Twenge JM (2011) ***Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics***. **Psychology of Aesthetics, Creativity and the Arts** 5: 200–207. doi: 10.1037/a0023195. Find this article online

…

"…the necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in *all the sciences* – arguably the key scientific theme of our times."

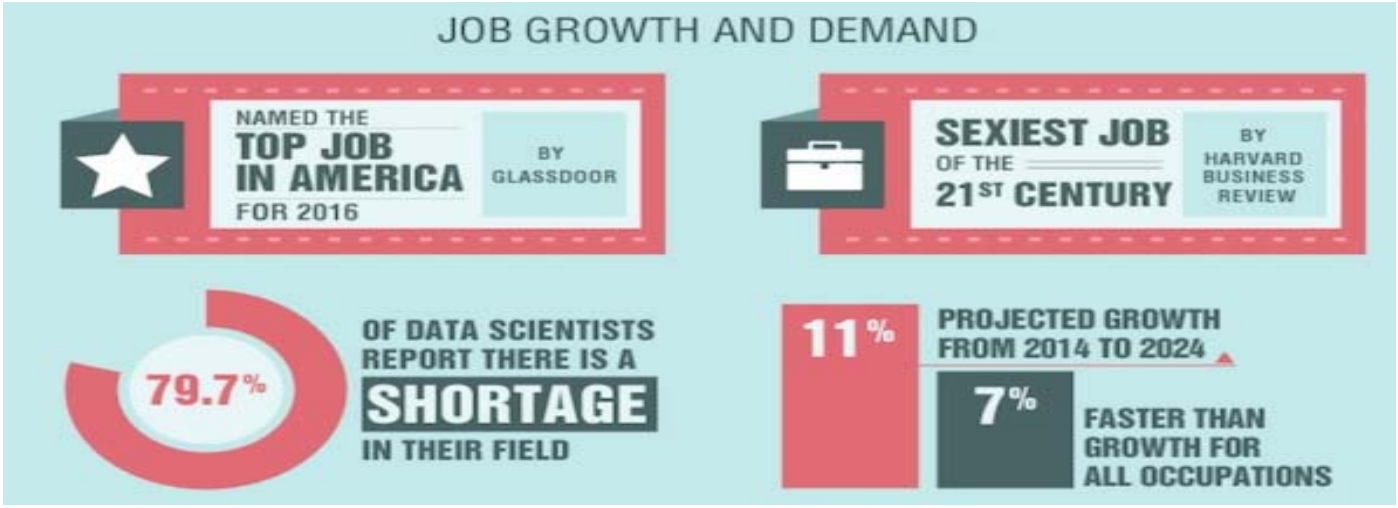
Francis X. Diebold

Paul F. and Warren S. Miller Professor of Economics

School of Arts and Sciences University of Pennsylvania

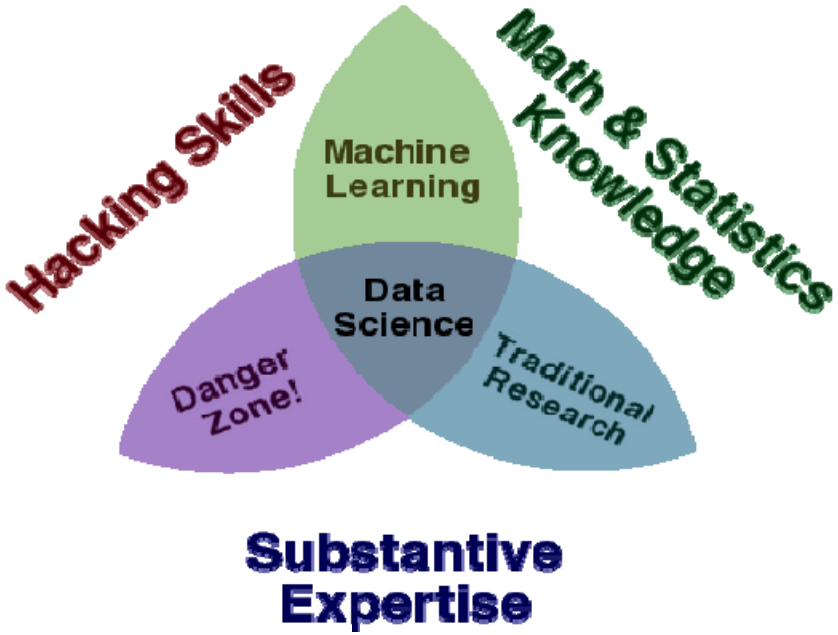
Big data demands

- KDnuggets report, 2017
  - Data scientist is selected as the sexiest job on 21st century by Harvard business review
  - From 2014 to 2024, the data scientist career path is expected to grow by 11%–14% faster than for all occupations.



http://www.kdnuggets.com/2017/05/data-science-need-to-know.html

Drew Conway's data science Venn diagram



- If you're a DBA, you need to learn to deal with unstructured data
- If you're a statistician, you need to learn to deal with data that does not fit in memory
- If you're a software engineer, you need to learn statistical modeling and how to communicate results.

New challenges in Big data

- Big data is not new?
- Very Large Database (VLDB) has been an important issue in research communities.
- Parallel processing has been a major research problem for the last century of computer science.
- **What are new challenges?**

IPA: Scalable and Parallelizable Processing of Influence Maximization for Large-Scale Social Network [ICDE 2013, best poster award]

1. **10x times faster** than PMIA (the state-of-the-art algorithm)

2. **Uses much less memory** than PMIA;

   - IPA successfully produces results on graphs of millions of nodes using 4GB memory where PMIA fails with 24GB memory.

3. **Accurately approximates influence spread;**

   - IPA's accuracy is close to that of Greedy solutions with 20k times MC simulation and is higher than that of PMIA overall.

4. **Can be applied to all IC-based models;**

   - PMIA cannot be applied to CT-IC model.

5. **Easily parallelized;**

   - The parallel IPA speeds up as # of CPU cores increases, and more speed-up is achieved for larger data sets.

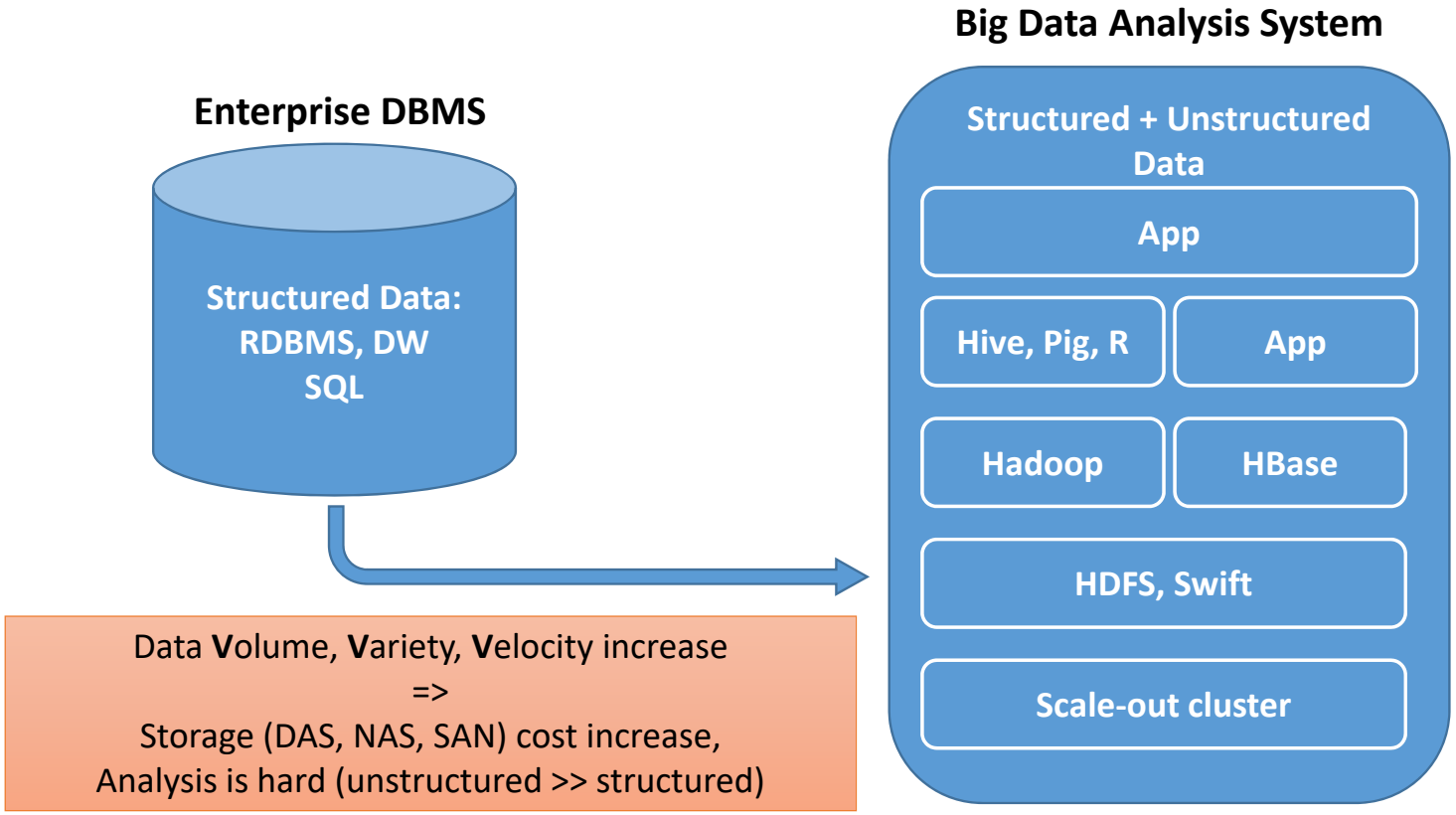
Scaling up to Billion-Nodes Network using Map-Reduce?

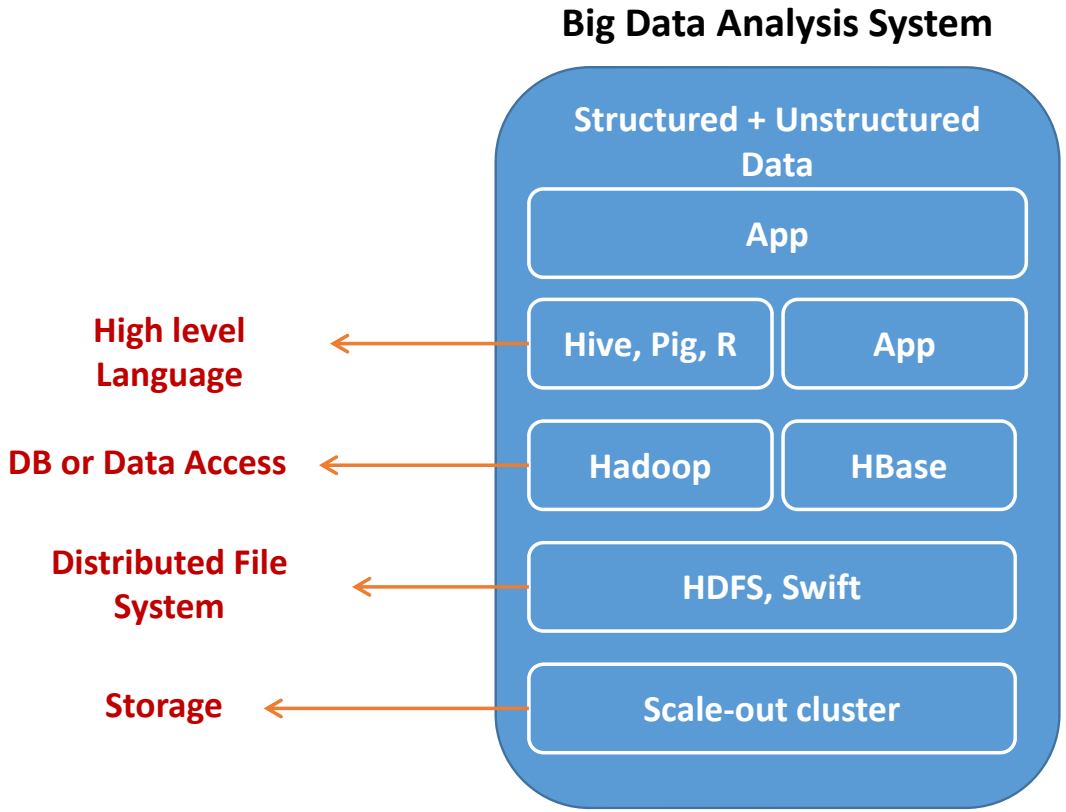
Very Hard !

Something is easily parallelized does NOT mean it can be easily "map-reduced".

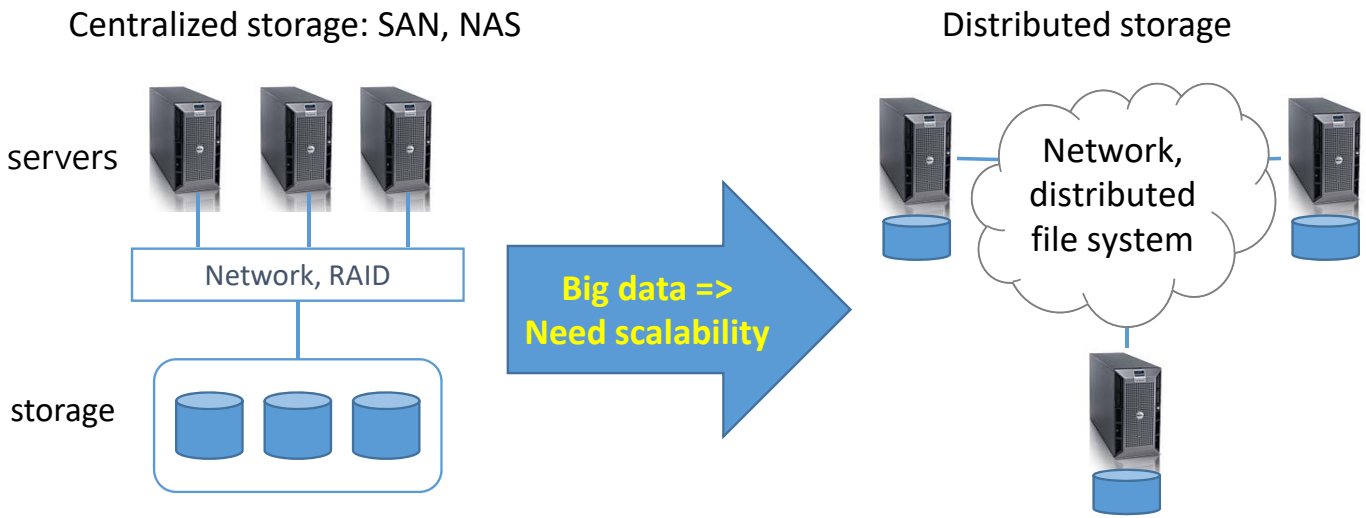
Big data processing ≠ Parallel data processing

How different?


Big Data Analysis System
Enterprise DBMS
Structured Data:
RDBMS, DW
SQL
Structured + Unstructured Data
App
Hive, Pig, R
App
Hadoop
HBase
HDFS, Swift
Scale-out cluster
Data Volume, Variety, Velocity increase
=>
Storage (DAS, NAS, SAN) cost increase,
Analysis is hard (unstructured >> structured)


Big Data Analysis System
Structured + Unstructured Data
App
Hive, Pig, R
App
Hadoop
HBase
HDFS, Swift
Scale-out cluster
High level Language
DB or Data Access
Distributed File System
Storage

Storage trend



- **Proprietary, Highly reliable HW**

=> Scale-up: Expensive

=> Fast data transfer

- **Commodity HW**

=> Scale-out: Inexpensive

=> Slow data transfer

=> Need new programming model !

Trend evolution

Data Trend(Big Data)

Storage Trend (Distributed) : Inexpensive Scale-out, but
Expensive Data Transfer!

Need New Programming Model to Minimize Data Transfer

Move operations instead of data!

MapReduce by Google

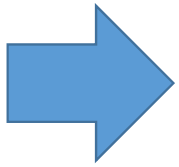
Hadoop and many subprojects

MapReduce Principles

- Run operation on data nodes: Move operations to Data
- Minimize data transfer

Programming is Hard!!!

A straightforward extension of parallel IPA algorithm produce *too many iterations* and *heavy data transfer from map to reduce*

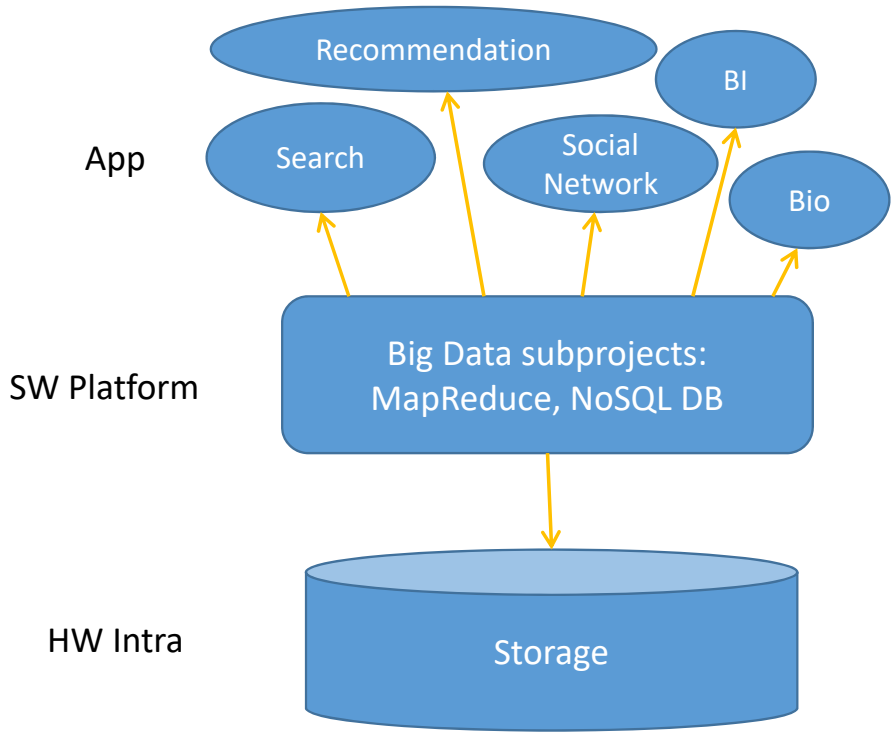


Design Tips

- Lower the work of reduce
  - Use combine if possible
- Compression of map's output helps decreasing network overhead
- Minimize iterations and broadcasting
  - Sharing information is minimized
- Use bulk reading
  - Too many invocation of map may incur too many function calls
- Design algorithm to have enough reduce functions
  - Having only a single reduce will not speed up

Big data subprojects

- Big data programming framework
  - MapReduce (Batch): HDFS & Hadoop, Dryad
  - MapReduce (Iterative): HaLoop, Twister
  - MapReduce (Streaming): Storm (Twitter), S4 (Yahoo), InfoSphere Streams (IBM), HStreaming
- NoSQL DB
  - HBase (Master, slaves), Cassandra (P2P, “Gossip”, no master server), Dynamo (Amazon), MongoDB (for text)
- Graph processing engine
  - Pregel, Giraph, Trinity, Neo4J, TurboGraph
- IoT platform
  - NoSQL DB + Analytics solutions
  - Allseen, Predix

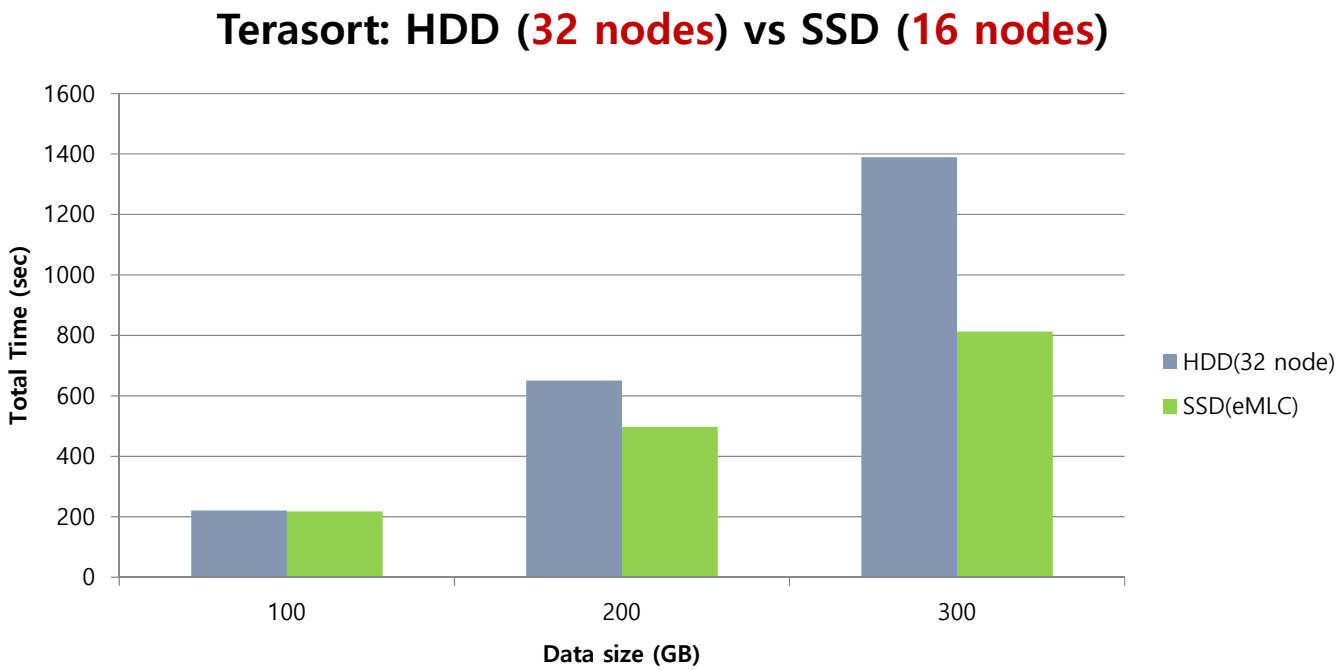


- Minimize Data Transfer
- Which platform?
- Generalization
- Feasible? Approximate?
- SSD-aware mining

- Move CPU to Data
- Minimize Data Transfer
- Search, Recommendation, ..
- Text, Graph, Multimedia, ..
- Batch, Streaming
- SSD-aware platform

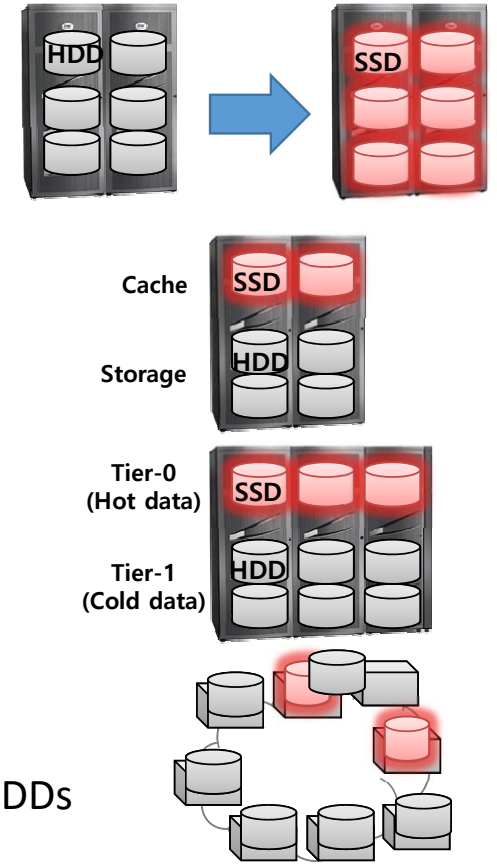
- Scalability
- Scale-out cost
- Energy efficiency
- Load balancing
- SSD where?

HDD vs SSD


Terasort: HDD (32 nodes) vs SSD (16 nodes)
1600
1400
1200
1000
800
600
400
200
0
Total Time (sec)
100
200
300
Data size (GB)
HDD(32 node)
SSD(eMLC)

SSD where to put?

- Replacement model
  - Replace HDD with SSD
  - Throw out HDD?
  - Big data => Expensive scale-out?
- Caching model
  - Use SSD as cache between memory and HDD
  - Ratio of SSD and HDD?
  - Data duplication?
- Tiering model
  - Put hot data to SSD and cold data to HDD
  - Data migration?
- Distributed model
  - Don’t care migration, don’t care ratio, no duplication, no need to throw out HDDs
  - Load balancing by Hadoop



Reality

- Linkedin
  - Develop NoSQL database “Voldemort” which uses SSDs
- Twitter
  - Optimize MySQL for SSDs (e.g., page-flushing behavior, reduction in writes to disk)
- Amazon
  - Develop SSD-based NoSQL database “DynamoDB” as a new service in AWS
- EBay
  - Replace its internal virtual storage layer with 100TB SSDs (2011)
  - Replace its internal virtual storage layer with 100TB SSDs (2011)
- Microsoft
  - Replace Bing Search runtime filesystem with Intel SSD (2011)
  - Uses Intel SSDs in their KeyValue storage for Bing social search
  - Microsoft Research is working on Flash Server Farm Called CORFU (Cluster Of Raw Flash Units)
- Facebook
  - Improve MySQL performance by adding Fusion-io as caching layer

Course objectives

- Big data is an interdisciplinary field involving multiple disciplines including databases, data mining, and machine learning. This course will serve as the first course for big data analytics which demands multiple skills not easy to obtain through conventional curricula.
- This course is designed for senior undergraduate or first-year graduate students in computer-related departments, who will do research in academia and also work in industry.

Course topics

- This course studies foundations of big data analytics and exercises using tools and languages. This course deals with two perspectives of big data — (1) system perspective, presenting how to store and process big data, and (2) methodology perspective, discussing how to build model from data.
- In the first half of the course, we study how to store, manage, search and analyze big data by utilizing popularly used solutions such as SQL, MapReduce, Hadoop, and Spark. In the second half, we study representative machine learning and data mining methods such as classification, clustering, recommender system, and link analysis.