Seminar
Big Data Analytics

Winter Term 2012/13

Lehrgebiet Informationssysteme

Weiping Qu
qu@cs.uni-kl.de

**AG Datenbanken und
Informationssysteme**

**AG Heterogene
Informationssysteme**

Goals

- Familiarize yourself with a scientific topic independently
- Find scientific literature on web or in the library, cite correclty (ACM digital library, DBLP, Citeseer, Google)
- Prepare a written composition, presentation, and discussion
- Time Management
- Don't copy and paste!

Written composition

Searching for Literature

- Some basic literature is provided by your supervisor

- You have to search for further literature

- Collect a list of related literature and send it to your supervisor

Annotated Table of Contents (TOC)

- Prepare an annotated TOC for your supervisor
(approx. 1.5 pages)

Written Composition

- LNCS Layout

- PDF Format required

- Length: 6000–8000 Words (net.) $\cong$ 15–20 Pages

- Correct and complete bibliography

Links

Homepage

➢ http://wwwlgis.informatik.uni-kl.de/cms/courses/seminar/

LNCS

➢ http://www.springer.com/computer/lncs?SGWID=0-164-6-793341-0

Typography

➢ http://www.zvisionwelt.de/typokurz.pdf

ACM Digital Library

➢ http://dl.acm.org/

DBLP

➢ http://www.informatik.uni-trier.de/~ley/db/

Presentation

- Length: 60 Minutes (45 minutes for the presentation, 15 minutes for the discussion)

- Presentation:
  – Projector (private or one of our notebooks)
  – Overhead slides

- You must submit your presentation electronically, one week after you gave your talk

- First talks are scheduled for
  – Friday, January 25th 2013
  – Room 36/336, 10:00 a.m – 12:00 p.m and 1:30 p.m – 4:30 p.m.

Big Data Analytics

- ➢ „Big data problem" occurs in Big Web companies, traditional enterprises, scientific data processing…

- ➢ Big data: large amount of data (petabyte), mostly unstructured, high scale-out..

- ➢ Analytics in enterprises:
  - ➢ data warehouses, parallel databases with novel hardwares and shared-nothing architecture
  - ➢ Relational, SQL interface

- ➢ Web companies (Google, Yahoo!, Amazon, Facebook..):
  - ➢ MapReduce, Hadoop, Key-value stores
  - ➢ Custom mapper and reducer functions or simple read/write data access.
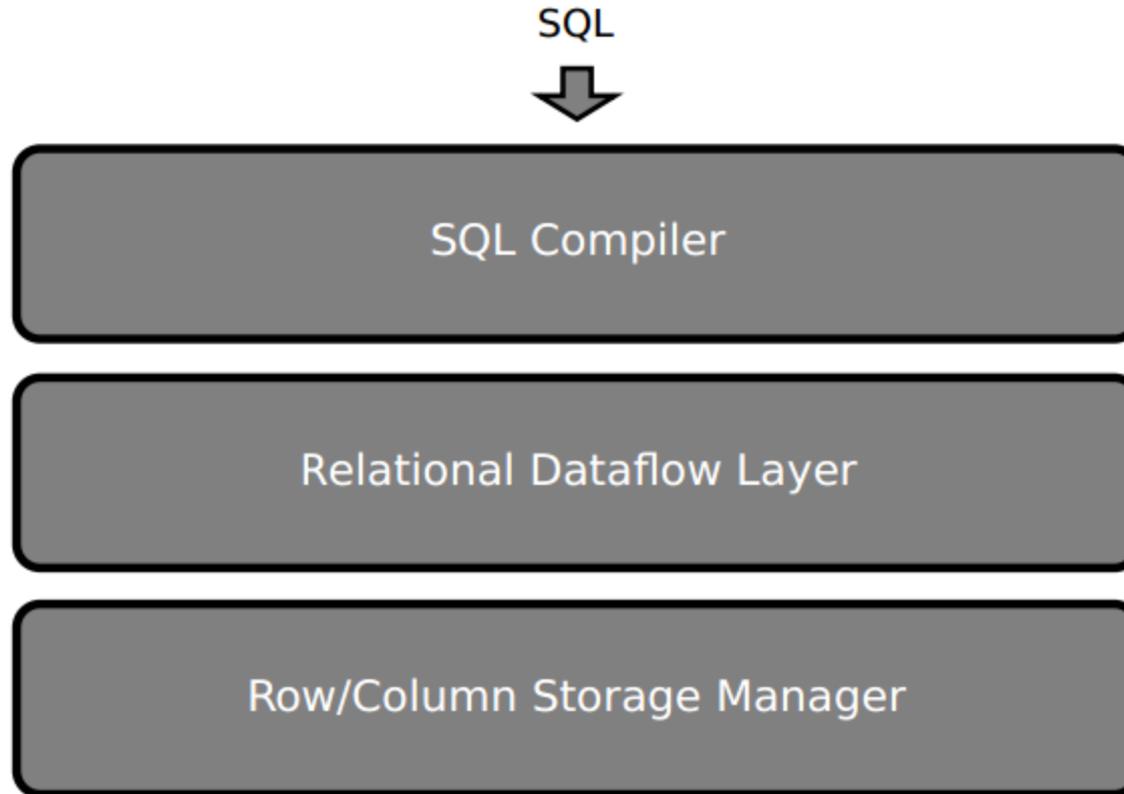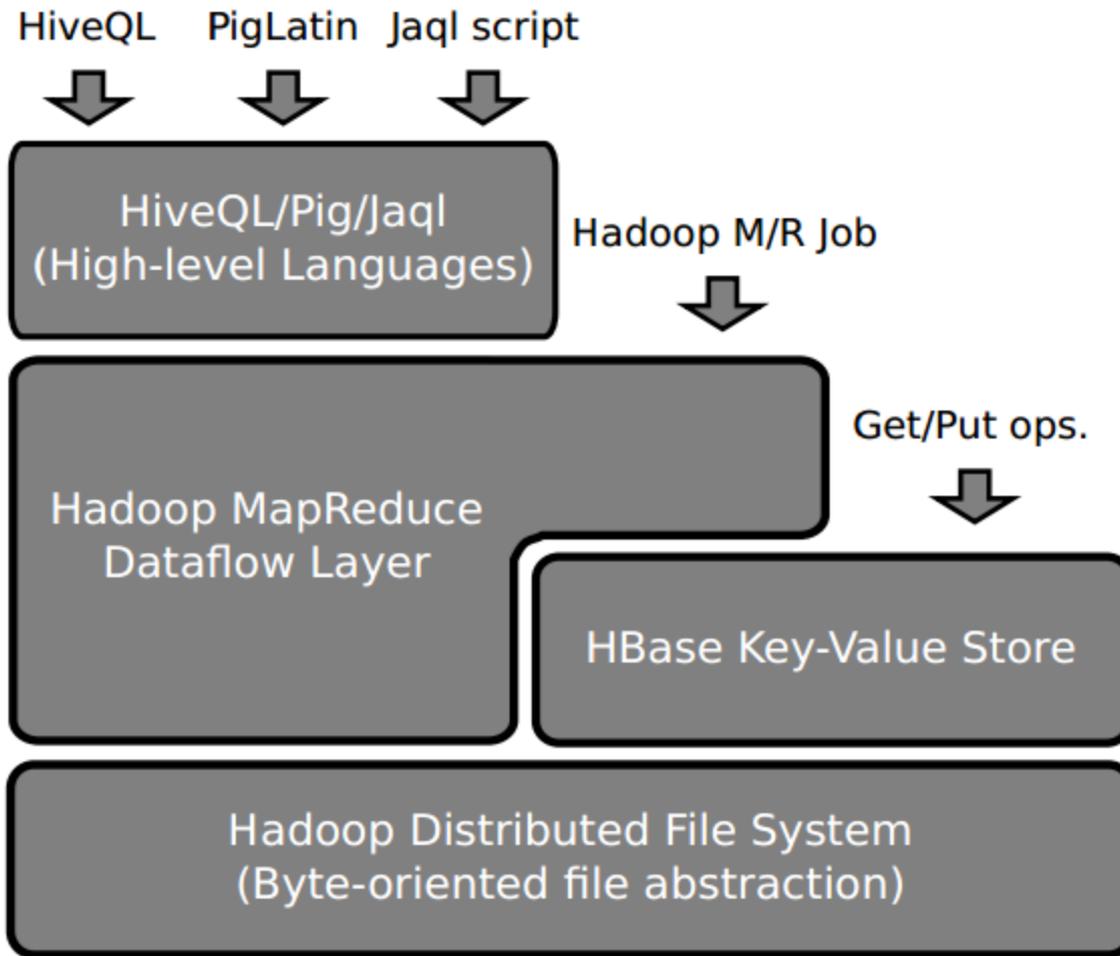
Parallel database software stack [1]

Topics

Topic 1:   Architectural design for scalable parallel computing

Topic 2:   Parallel computing systems

Topic 3:   Data flow system and operator model

Topic 4:   NoSQL databases

Topic 5:   Hadoop++

Topic 6:   Integrating MapReduce and relational databases

Topic 7:   Large-scale graph processing

Topic 8:   Database in the cloud

Topic 9:   In-memory databases

Topic 10:   Scientific data processing

How you successfully pass the seminar

- **"Unbenoteter Schein":**
  - Meet the deadlines!
  - Decent presentation
  - Be present when others give their talks
- **"Benoteter Schein"**
  - See above
  - **Criteria for grading:**
    - Quality of your written composition
    - Quality of your presentation (including your slides)
    - Discussion
    - Meeting the deadlines
    - Overall impression of your supervisor
    - ...

- Friday, October 19th: Kick-off meeting
- Friday, November 2nd: Deadline for literature list
- Friday, November 16th: Deadline for annotated TOC
- Monday, January 14th: Deadline written composition
- Friday, January 25th: Presentations and
   Monday, January 28th: Presentations

All deadlines are strict!

Questions

Select your topic...now!

Topic 1:   Architectural design for scalable parallel computing

Topic 2:   Parallel computing systems

Topic 3:   Data flow system and operator model

Topic 4:   NoSQL databases

Topic 5:   Hadoop++

Topic 6:   Integrating MapReduce and relational databases

Topic 7:   Large-scale graph processing

Topic 8:   Database in the cloud

Topic 9:   In-memory databases

Topic 10:   Scientific data processing

[1] Vinayak Borkar, Michael J. Carey, Chen Li *Inside "Big Data Management": Ogres, Onions, or Parfaits?* In EDBT, 2012